Université de Sherbrooke

# Effet du contexte nucléotidique dans la prédiction, le repliement et la fonction des structures G-quadruplex d'ARN

Par
Rachel Jodoin
Programme de Doctorat en biochimie

Thèse présentée à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade de philosophiae doctor (Ph. D.)
en Biochimie

Sherbrooke, Québec, Canada
Février 2019

Membres du jury d'évaluation
Pr Jean-Pierre Perreault, Ph. D. – Département de biochimie
Pr Martin Bisaillon, Ph. D. – Département de biochimie
Pr Éric Massé, Ph. D. – Département de biochimie
Dr Maxime Richer, M.D., Ph. D. – Départements de pathologie et de biochimie
Pr François Boudreau, Ph. D. – Département d'anatomie et de biologie cellulaire
Pr Jerry Pelletier, Ph. D. – Department of Biochemistry, McGill University

# RÉSUMÉ

**Effet du contexte nucléotidique dans la prédiction, le repliement et la fonction des structures G-quadruplex d'ARN**

Par
Rachel Jodoin
Programme de Doctorat en biochimie

Thèse présentée à la Faculté de médecine et des sciences de la santé en vue de l'obtention du diplôme de philosophiae doctor (Ph. D.) en biochimie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

L'ARN est une biomolécule essentielle dont la fonction est étroitement reliée à sa structure secondaire. Les G-quadruplexes (G4) formés de l'empilement de tétrades de G, qui sont des agencements coplanaires de guanines stabilisées par des paires de bases Hoogsteen et la présence de cations potassium, sont des structures secondaires très stables. Les G4 sont adoptés par les séquences respectant le motif canonique suivant : $(G_3\text{-}N_{1\text{-}7})_3G_3$. Les G4 d'ARN (rG4) présents dans les ARN messagers (ARNm) sont connus pour réguler de nombreux processus post-transcriptionnels tels que l'épissage et la traduction.

Cependant, plusieurs séquences respectant le motif canonique rG4 ne forment pas la structure lorsqu'elles sont testées expérimentalement. À l'inverse, plusieurs séquences qui divergent du motif ont été démontrées pour adopter la structure rG4. Cela suggère que d'autres facteurs influencent le repliement. La présence de cytosines dans le contexte nucléotidique rapproché du motif rG4, qui pourraient former des paires de bases G-C avec les séries de G, a été proposée comme un facteur compétitif à la formation de rG4. Afin de confirmer cette hypothèse, une technique *in vitro* permettant d'analyser le repliement rG4 de plus longues séquences était nécessaire. La cartographie *in-line* a été adaptée puis appliquée sur des séquences rG4 potentielles avec un large contexte nucléotidique. Cela a permis de confirmer l'influence des séries de G et des séries de C dans l'environnement immédiat sur le repliement du rG4. Un nouveau score prédictif tenant compte de ces séries a été établi permettant d'améliorer la sensibilité et la spécificité des prédictions de repliement rG4.

Les outils d'expérimentation et de prédiction développés dans le cadre des travaux ayant mené à cette thèse ont permis de prédire et d'évaluer le repliement de plusieurs candidats rG4 de motifs variés présents dans les 5'UTR d'ARNm. De plus, la présence enrichie de rG4 dans des ARNm associés au cancer colorectal a pu être détectée. Le rG4 irrégulier identifié dans le transcrit BAG-1 et son contexte nucléotidique complet en 5'UTR ont été étudiés en détail afin de mesurer l'effet du rG4 sur la structure secondaire globale. Les effets du rG4 sur la traduction ainsi que son interaction fonctionnelle avec de nombreux éléments régulateurs de la traduction situés en *cis,* tels des codons de départ alternatifs, un uORF et une structure IRES ont été mesurés. La considération du contexte nucléotidique est essentielle afin de correctement prédire et de permettre le repliement rG4, ainsi que pour déterminer ses fonctions biologiques. **Mots-clés** : G-quadruplex, ARN messager, 5'UTR, Structure secondaire, Prédiction, Cartographie, Traduction, Cancer.

# SUMMARY

**Effect of the surrounding nucleotide context on the prediction, the folding and the function of RNA G-quadruplex structures**

By
Rachel Jodoin
Biochemistry Doctoral Program

Thesis presented at the Faculty of Medicine and Health Sciences for the obtention of Doctor's degree diploma philosophiae doctor (Ph.D.) in Biochemistry, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

RNA is an essential biomolecule whose function is highly related to its secondary structure. The G-quadruplexes (G4s) formed by the stacking of G-quartets, which are co-planar arrangements of guanines stabilised by Hoogsteen base-pairs and potassium cations, are highly stable secondary structures. G4s are adopted by sequences corresponding to the consensus motif: $(G_3\text{-}N_{1\text{-}7})_3G_3$. RNA G4s (rG4s) located in messenger ARNs (mRNAs) are known to regulate numerous post-transcriptonal processes such as splicing and translation.

However, many sequences presenting the consensus rG4 motif do not fold into the structure when they are experimentally validated. In contrast, many sequences differing from the motif were demonstrated to adopt the rG4 structure. This suggest that other factors influence the folding. It was proposed that the presence of many cytosines in the proximal nucleotide context of the rG4 motif was a competing factor to rG4 folding by forming G-C base-pairs with the G-tracts. To confirm this hypothesis, an *in vitro* technique allowing to probe the rG4 folding of longer sequences was necessary. The in-line probing technique was adapted and applied to potential rG4 sequences candidates with their large nucleotide context. These in-line probing results confirmed the influence of the G- and C-tracts of the surrounding environment on the rG4 folding. A new predictive score calculating those tracts was established which improved the sensibility and the specificity of the rG4 folding predictions.

The experimentation and scoring tools developed as part of the work leading up to this thesis allowed to predict and to probe the secondary structures of many rG4 candidates with various motifs located in the 5'UTR of mRNAs. Furthermore, the enrichment of rG4 in mRNAs associated to colorectal cancer was detected. The irregular rG4 identified in the BAG-1 transcript and its complete 5'UTR nucleotide context were studied in detail in order to measure the rG4 effect on the global secondary structure. The rG4 effects on translation as well as its functional interaction with several *cis* translational regulatory elements such as alternative start codons, an uORF and an IRES structure was measured. The analysis of the nucleotide context is essential to predict accurately if it allows rG4 folding, as well as to decipher the biological functions of the structure.

**Keywords**: G-quadruplex, messenger RNA, 5'UTR, Secondary structures, Prediction, Probing, Translation, Cancer.

# TABLE DES MATIÈRES

# LISTE DES FIGURES

# LISTE DES TABLEAUX

# LISTE DES ABRÉVIATIONS

| | |
|---|---|
| 2'-OH | 2'-hydroxyl |
| $^{32}P$ | Phosphore-32 |
| A | Adénine |
| ADN | Acide désoxyribonucléique |
| ARN | Acide ribonucléique |
| ARNm | ARN messager |
| ARNt | ARN de transfert |
| ASO | Oligonucléotide antisens / *Antisense oligonucleotide* |
| ATP | Adénosine triphosphate |
| AUC | Aire sous la courbe */ Area under the curve* |
| BAG-1 | *BCL2-associated athanogene 1* |
| BBP / BPB | Bleu de bromophénol / *Bromophenol blue* |
| Bp / pb | *Base-pair* / Paire de base |
| BzCN | Benzoyl cyanide |
| C | Cytosine |
| C-terminale | Carboxy-terminale |
| cC | Cytosines consécutives / *concecutive Cytosines* |
| CD / DC | Dichroïsme circulaire / *Circular dichroism* |
| cDNA / ADNc | ADN complémentaire / *Complementary DNA* |
| CDS | *Coding sequence* |
| cG | Guanines consécutives / *consecutive Guanines* |
| cG/cC | Score G consécutifs sur C consécutifs |
| cPDS | Carboxypyridostatine |
| cpm | Compte-par-minute / *Count-per-minute* |
| CRC | Cancer colorectal / *Colorectal cancer* |
| ddNTP | Didésoxyribonucleotide |
| ddPCR | *Digital droplet PCR* |
| dNTP | Désoxyribonucleotide |
| DMS | Dimethylsulfate |
| DMSO | Diméthylsulfoxyde |
| DTT | Dithiothréitol |
| EDTA | Acide éthylènediaminetétraacétique |
| EMSA | *Electropohretic mobility shift assay* |
| Fluc | *Firefly luciferase* |
| FRET | *Förster resonance energy transfer* |
| G | Guanine |
| G4 | G-quadruplex |
| G4H | G4Hunter |
| G4NN | *G4RNA Neural Network* |
| GO | *Gene-ontology* |
| HCV | Virus de l'hépatite C / *Hepatitis C virus* |
| HRP | *Horseradish peroxydase* |

| | |
|---|---|
| Hsp70/Hsc70 | *Heat shock protein 70, Heat-shock chaperone 70* |
| HEPES | Acide 4-(2-hydroxyéthyl)-1-pipérazine éthane sulfonique |
| IRES | Site d'entrée interne du ribosome / *Internal ribosome entry site* |
| $K^+$ | Ion potassium |
| $Li^+$ | Ion lithium |
| LNA | *Locked nucleic acids* |
| lncARN | Long ARN non codant |
| $M^+$ | Cation monovalent |
| $m^7G$ | 7-methylguanylate |
| mfe, MFE | Énergie libre minimale / *Minimum free energy* |
| miARN | Micro ARN |
| N | Nucléotide : A, C, G, T, U |
| N-terminal | Amino-terminale |
| $Na^+$ | Ion Sodium |
| NLS | *Nuclear localisation signal* |
| NMM | N-Méthyl-mésoporphyrine IX |
| NMR / RMN | Résonance magnétique nucléaire */ Nuclear magnetic resonance* |
| nt | Nucléotide |
| NTP | Nucléotide tri-phosphate |
| ORF | *Open Reading Frame* |
| PAGE | *Polyacrylamide gel electrophoresis* |
| pAS | Signal de polyadénylation / *Polyadenylation signal* |
| PEG | Polyethyleneglycol |
| PG4 | G-quadruplex potentiel/ *Potential G4* |
| PI3-Kinase | Phosphoinositide-3-kinase |
| PIC | *Pre-initiation complex* |
| PQS | *Probable quadruplex sequence* |
| PVDF | *Polyvinylidene difluoride* |
| QGRS | *Quadruplex forming G-Rich Sequences* |
| RBP | *RNA binding protein* |
| RBS | *Ribosome binding site* |
| rG4 | G-quadruplex d'ARN |
| RNase | Ribonucléase |
| RPF | *Ribosome-protected fragments* |
| RIPA | *Radioimmunoprecipitation assay* |
| Rluc | *Renilla luciferase* |
| rNTP | Ribonucléotide |
| ROC | *Receiver-operator characteristic* |
| TmPyP4 | Meso-5,10,15,20-Tetrakis-(N-methyl-4-pyridyl)porphine |
| RT | Transcriptase ou Transcription inverse / *Reverse Transcriptase or transcription* |
| RTS | *Reverse transcriptase stalling assay* |
| SAFA | *Semi-Automated Footprinting Analysis* |
| SHAPE | *Selective 2'Hydroxyl Acylation analysed by Primer Extension* |
| SDS | Sodium dodecyl sulfate |
| siARN | Petit ARN interférent / *Small interfering RNA* |
| T | Thymine |

| | |
|---|---|
| TERRA | *Telomeric repeat-containing RNA* |
| $T_m$ | Température de fusion / *Melting temperature* |
| U | Uracile |
| UBL | *Ubiquitin binding ligand* |
| uORF | *Upstream ORF* |
| UTR | Région non traduite / *untranslated region* |
| XC | Xylène cyanol |

# INTRODUCTION

Le dogme de la biologie moléculaire débute par la double hélice d'acide désoxyribonucléique (ADN) qui est le support stable et réplicable pour l'ensemble du code génétique (Watson et Crick, 1953). L'utilisation de cette information encodée sous forme de suite nucléotidique débute par la transcription. Cette étape correspond à la copie d'un segment de ce code, appelé gène, sous la forme d'un intermédiaire d'acide ribonucléique (ARN) appelé ARN messager (ARNm). Par la suite, l'ARNm peut sortir du noyau pour rejoindre le cytoplasme où il sera reconnu par la machinerie de synthèse protéique pour que l'information de la suite de triplets nucléotidiques soit décodée et traduite en séquences d'acides aminés qui seront assemblés en peptides et finalement en protéines, les unités fonctionnelles dans la cellule (**Figure 1**)(Crick, 1970). Chacune des étapes de cette suite est hautement régulée pour maintenir la fonction tissulaire et l'homéostasie de la cellule et ainsi répondre à l'ensemble des besoins métaboliques d'un organisme vivant (Watson *et al.*, 2009).

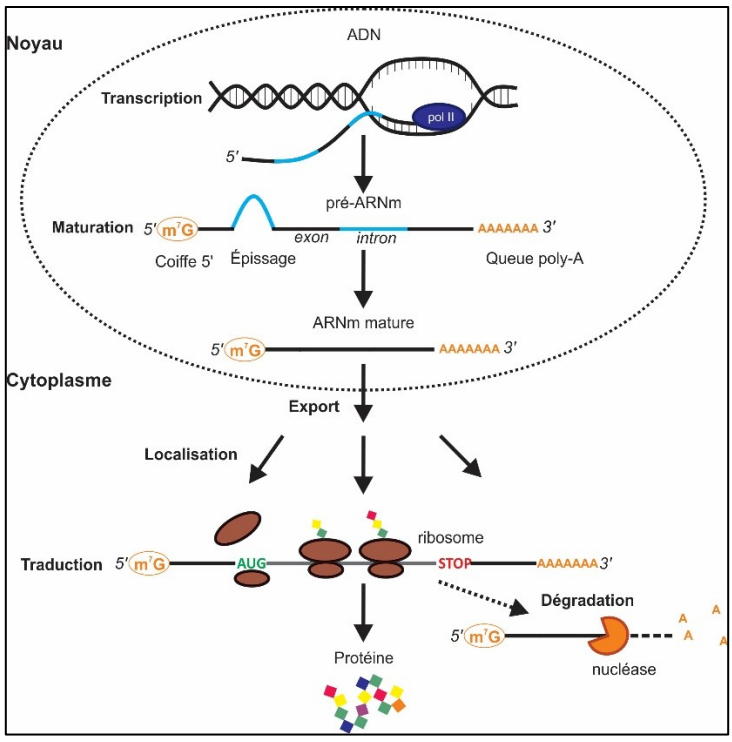



**Figure 1** – Dogme de la biologie moléculaire.

L'information génétique est stockée au niveau du noyau sous forme d'ADN. Afin d'utiliser cette information, l'ADN est transcrit par une polymérase en pré-ARN messager. Suite à sa maturation,

qui consiste en l'épissage des introns et en l'ajout de la coiffe 5' et de la queue poly-A, le transcrit mature sera exporté au cytoplasme et transporté jusqu'à sa localisation dans la cellule afin que sa séquence soit traduite en chaîne peptidique par les ribosomes pour former la protéine, l'unité fonctionnelle. Ultimement, le brin d'ARNm sera dégradé après avoir été traduit et avoir épuisé son temps de vie.

Dans cette description, l'ARN semble jouer un rôle passif de simple support temporaire d'information, mais il n'en est rien. Par leur grande versatilité, qui s'explique par la présence du ribose dans leur squelette phosphosdiester et leur forme simple-brin, les molécules d'ARN sont présentes en multiples familles qui possèdent une panoplie de fonctions cellulaires. L'ARN possède les avantages des deux mondes : tout comme l'ADN, cette molécule peut encoder de l'information et former des appariements de bases et, tout comme les protéines, elle peut aussi posséder des fonctions régulatrices et catalytiques.

### Séquences primaires, structures secondaires et tertiaires

L'émergence de ces fonctions découle des trois niveaux successifs d'organisation de la molécule. Le premier niveau est la séquence primaire, la suite des différents nucléotides (nt) adénine (A), uracile (U), cytosine (C) et guanine (G), transcrite du génome d'ADN pour former le transcriptome d'ARN. Cette séquence forme les codons des ARNm qui dicteront la synthèse protéique. C'est également cette séquence primaire qui permettra les appariements complémentaires avec d'autres ARN comme le font les petits ARN interférents (siARN) ou les microARNs (miARN).

Le second niveau d'organisation de l'information est la structure secondaire, c'est-à-dire la structure adoptée par la formation intramoléculaire d'appariements de nucléotides, qui découle donc de sa séquence primaire. Les structures secondaires pouvant être adoptées par les ARN sont des tiges double-brin (*stem)* qui sont reliées entre elles et à leurs extrémités par des séries de nucléotides non appariés appelés boucles, ce qui permet de former des tiges-boucles (*stem-loop* ou *hairpin*) ou des multi-boucles où plusieurs tiges s'attachent. Ces tiges peuvent être parfaitement appariées, ou être interrompues par un nucléotide non apparié, appelé renflement (*bulge*) ou par de plus longs segments de nucléotides libres qui forment des boucles internes (**Figure 2A, B**). Ces structures secondaires possèdent différentes stabilités, qui sont proportionnelles au nombre et au type d'appariements. Par exemple, un appariement d'une paire de bases (pb) G-C est plus stable que les paires de bases G-U et A-U. L'assemblage de ces différentes structures secondaires peut aussi former des motifs, ou

mettre en évidence des segments de séquences (dans les boucles par exemple) qui sont reconnus par des facteurs *trans,* soit des protéines ou d'autres ARN qui viennent s'apparier.

Cela nous emmène au troisième niveau d'organisation, la structure tertiaire. Celle-ci émerge des appariements et liens formés entre des structures secondaires distantes. La structure tertiaire est donc obtenue par le repliement sur elles-mêmes des diverses structures secondaires pour former des pseudonœuds, des *kissing loops*, des triplex, etc. (**Figure 2C**) Cela entraîne aussi la formation de jonctions diverses entre ces structures réunies entre autres par des *A-turn, kink turn,* etc. Cette organisation en trois dimensions peut aussi servir de motif de reconnaissance pour des partenaires protéiques et rapprocher des éléments distants et permettre ainsi l'obtention de propriétés catalytiques telle que le clivage ou encore la ligation (Cruz et Westhof, 2009).

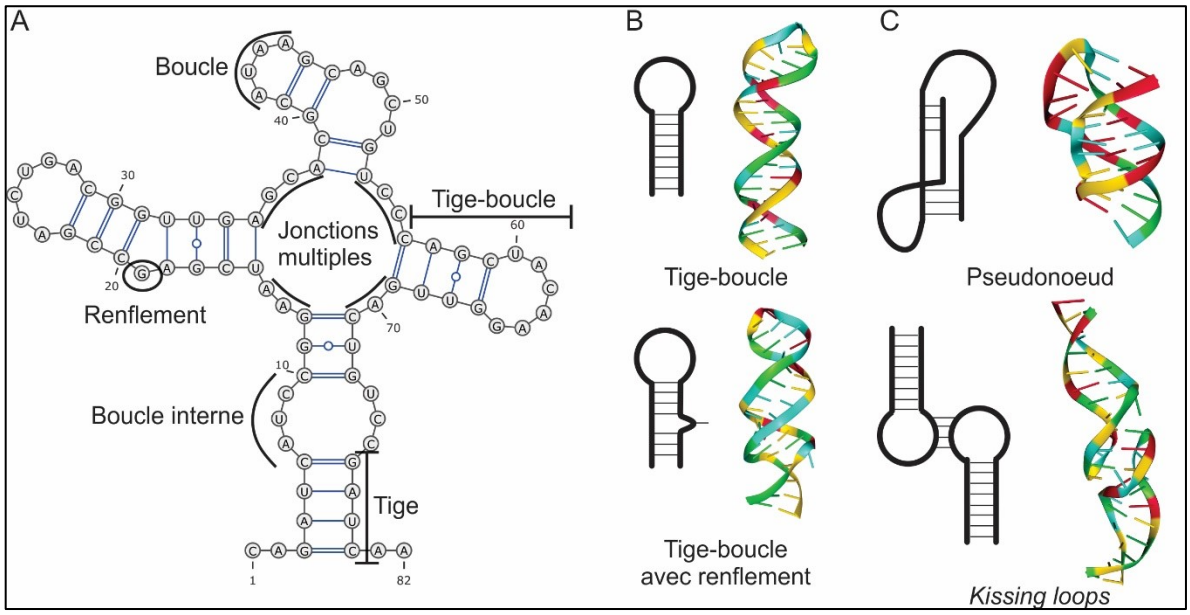



**Figure 2** – Structures secondaires et tertiaires de l'ARN.

(A) Séquence d'ARN adoptant une structure secondaire. Les différents types de structure secondaires sont indiqués. (B) Représentation schématique et 3D d'une tige-boucle PDB : 2L2J, (Stefl *et al.*, 2010) et d'une tige-boucle avec renflement, PDB : 1TXS, (Du *et al.*, 2004). (C) Représentation de deux exemples de structures tertiaires, un pseudonoeud et une paire de *kissing loops* PDB : 5KMZ et 2MIO, respectivement (Bouchard et Legault, 2014 ; Jiang *et al.*, 2015). Les couleurs bleu pâle, rouge, jaune et vert représentent respectivement les nucléotides U, A, C et G.

De ces différents niveaux d'organisations structurales découle la fonction des ARN. Ce concept essentiel de « structure-fonction » permet d'expliquer les fonctions de plusieurs ARN. Pour nommer quelques exemples, il y a les ribozymes, ces petits ARN avec des

fonctions catalytiques ; les riborégulateurs en bactérie (*riboswitches)* qui en étant liés par un ligand changent leur structure secondaire, ce qui résulte en une fonction régulatrice d'inhibition ou d'activation de l'expression des gènes ; les ARN de transfert (ARNt) qui permettent de décoder les ARNm, d'apporter l'acide aminé nécessaire et qui sont reconnus par le ribosome, cette machinerie de synthèse qui forme le lien peptidique et qui est elle-même composée d'ARN ribosomaux essentiels à sa fonction, et il y en a plusieurs autres.

La connaissance et la prédiction des structures adoptées par les ARN sont donc les bases permettant d'identifier leurs stabilités et leurs fonctions biologiques. Cela permet de comprendre comment ces fonctions émergent et de déterminer comment elles sont régulées par la cellule ou modulées par des agents externes (Mortimer *et al.*, 2014).

### Les structures G-quadruplexes, fondements chimiques et atomiques

Les G-quadruplexes (G4) sont des structures secondaires d'acides nucléiques, pouvant être formées autant par des séquences d'ADN que d'ARN. Comme le nom l'indique, cette structure est adoptée par des séquences riches en bases azotées guanines (G). Les cinq bases azotées : guanine (G), cytosine (C), adénine (A), thymine (T) et uracile (U) possèdent chacune trois faces possibles d'interaction : l'interface vers le sucre (ribose pour l'ARN ou désoxyribose pour l'ADN), l'interface canonique Watson-Crick et l'interface Hoogsteen (**Figure 3A, B**). Les paires de bases purine-pyrimidine G-C, A-T et A-U formées par des ponts hydrogène entre les interfaces Watson-Crick des deux bases azotées constituent les structures secondaires canoniques. Les structures G4 sont dites non canoniques. L'unité de base des G4 est la tétrade : un appariement coplanaire de 4 guanines reliées entre elles par des paires de bases formées à l'interface Hoogsteen de la base azotée (**Figure 3C**). Comparativement à l'appariement Watson-Crick canonique G-C qui est formé de 3 ponts hydrogène, la tétrade comporte 8 ponts hydrogène. Ceux-ci sont formés entre les atomes donneurs aux positions N1 et N2 d'une guanine avec les atomes receveurs aux positions O6 et N7 de la guanine adjacente, ce qui forme 4 ponts N1-O6 et 4 ponts N2-N7 (Gellert *et al.*, 1962). La surface de la tétrade est donc deux fois plus grande que la surface d'une paire de bases G-C (Neidle, 2012). Dans cette disposition des bases en tétrade, les atomes d'oxygène chargés négativement du groupement carboxyle (atome de carbone relié par une double liaison à un oxygène) de chaque guanine se retrouvent orientés vers le centre de la tétrade.

Pour éviter la répulsion électronique et assurer la stabilité de la tétrade, cette charge partielle négative doit être contrebalancée. Pour ce faire, un cation monovalent ($M^+$, chargé positivement) doit se coordonner au centre de la cavité, entre les plans formés par deux tétrades (**Figure 3C**).

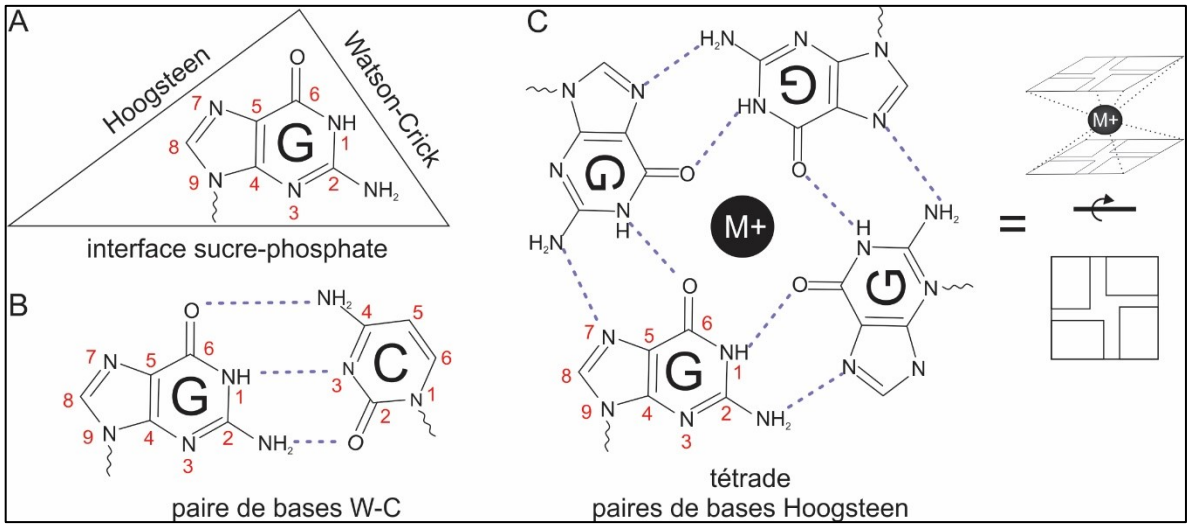



**Figure 3** – Base azotée guanine et tétrade.

(A) Représentation des 3 faces d'interaction de la base azotée guanine. (B) Appariement canonique Watson-Crick d'une paire de bases G-C. (C) Tétrade de G formée d'appariements Hoogsteen, stabilisée par la coordination d'un cation monovalent ($M^+$) en son centre et sa représentation schématique. La numérotation en rouge indique les positions des atomes de la base azotée.

L'empilement Pi (π) successif (*π-stacking*) de plusieurs tétrades, ayant chacune une rotation de 31° par rapport à l'autre, forme la structure d'hélice à 4 brins tournant vers la droite, d'où l'appellation quadruplex (**Figure 4A**). Ces tétrades sont reliées entre elles par le squelette phosphodiester de l'acide nucléique. Le squelette est perpendiculaire au plan de la tétrade. Ainsi, chaque guanine consécutive d'une séquence fera partie de tétrades différentes, empilées une par-dessus l'autre (**Figure 4B**). Les cations monovalents se retrouvent coordonnés entre les plans formés par les tétrades, dans le « canal » se formant au centre de chacune des tétrades (**Figure 3C et 4A, C**). Tout comme une hélice double-brin, cette hélice quadruple-brin présente à sa surface des sillons (*groove*) qui sont formés par le squelette phosphodiester. Les boucles reliant les tétrades peuvent s'insérer dans ces sillons.

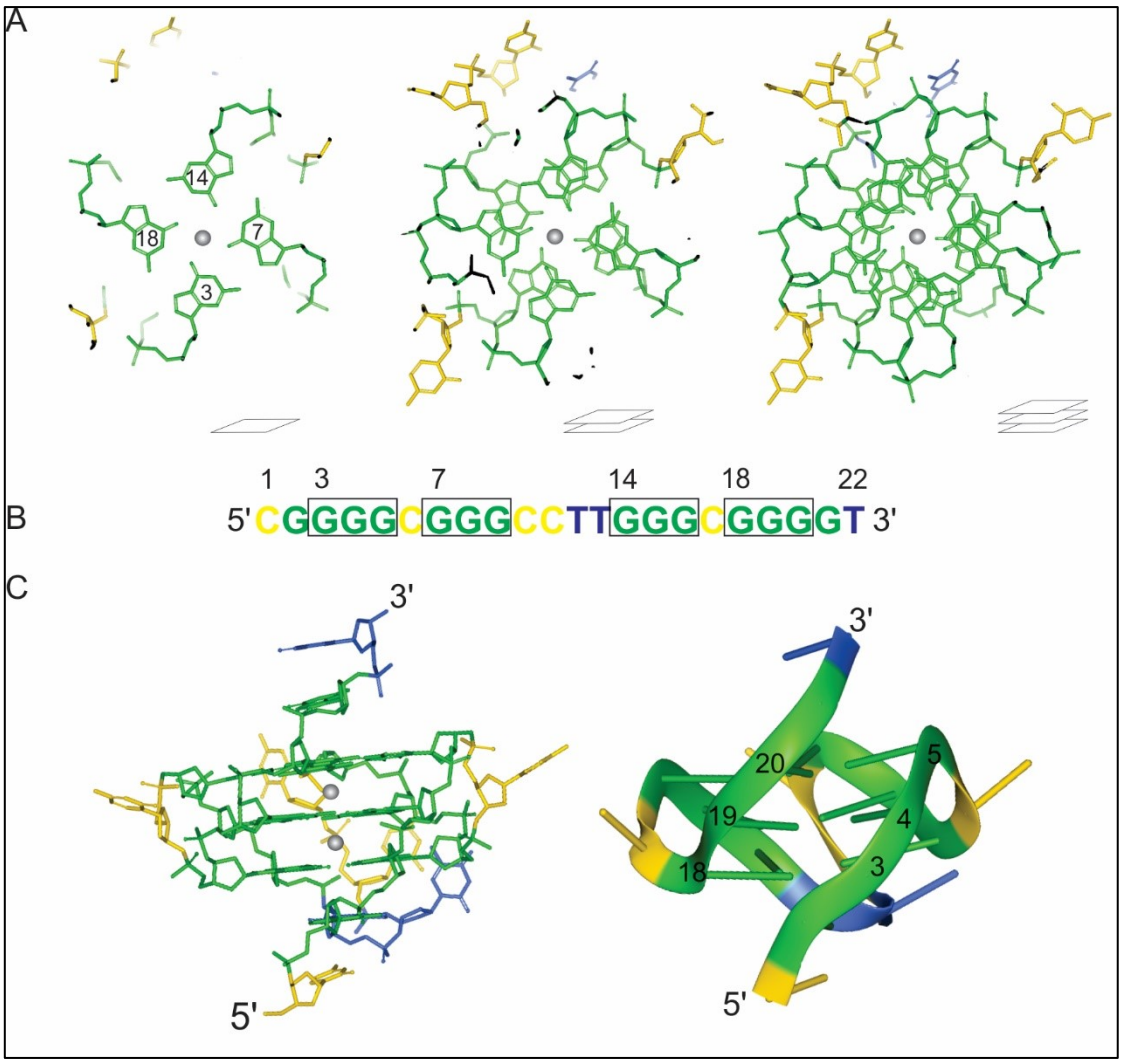



**Figure 4** – Empilement des tétrades et structure G4.

(A) Représentation atomique de l'empilement successif de 3 tétrades, vue en plongée. (B) Séquence du G4 situé dans le promoteur du gène VEGF, les séries de G participants aux tétrades sont encadrées. (C) Représentation atomique et représentation du squelette phosphodiester en ruban de la structure en hélice du cristal du G4 intramoléculaire parallèle du promoteur du gène VEGF [pdb : 2M27, (Agrawal *et al.*, 2013)], vue latérale. L'orientation 5'-3' du brin est indiquée ainsi que les positions des quelques guanines participant aux tétrades.

### Molécularité et topologie des G4

Les deux facteurs essentiels à la formation de G4 sont la formation des tétrades et leur empilement (**Figure 5A**). Donc, le strict minimum requis est la présence de quatre séquences composées d'au moins deux guanines consécutives qui formeront les « arêtes » du G4. Plusieurs combinaisons moléculaires sont possibles pour arriver à cette fin. Les G4 peuvent être intermoléculaires : plusieurs brins d'acides nucléiques distincts fournissant les séries de

G consécutifs. Ils peuvent donc être tétra-moléculaires, tri-moléculaires ou bi-moléculaires. Afin d'obtenir un G4 unimoléculaire et donc intramoléculaire, la séquence du brin doit contenir les quatre séries de deux G consécutifs minimaux (**Figure 5B**). Les nucléotides intercalants, qui séparent les séries de G impliquées dans les tétrades, forment les boucles. Un G4 intramoléculaire possédera trois boucles pour relier les séries de G. Ces boucles peuvent être composées de tous les nucléotides (A, C, G, T et U).

Les G4 peuvent adopter plusieurs topologies différentes, qui sont définies par l'orientation relative entre elles des séries de G formant les tétrades. L'orientation est dite parallèle lorsque le sens 5'-3' du squelette phosphodiester des séries de G est identique pour les quatre arêtes du G4 et antiparallèle lorsque deux brins adjacents sont en sens opposé. Une topologie mixte où certains brins sont parallèles entre eux et d'autres antiparallèles est aussi possible (**Figure 5C**).

Selon les différentes orientations et la molécularité du G4, différents types de boucles reliant les séries de G sont possibles. Les boucles diagonales relient des séries de G opposées. Les boucles latérales permettent de relier des arêtes adjacentes qui sont antiparallèles. Pour relier des arêtes adjacentes parallèles, les boucles formées sont de type « hélice » (aussi appelées *chain-reversal* ou *propeller*) (**Figure 5B**) (Burge *et al.*, 2006).

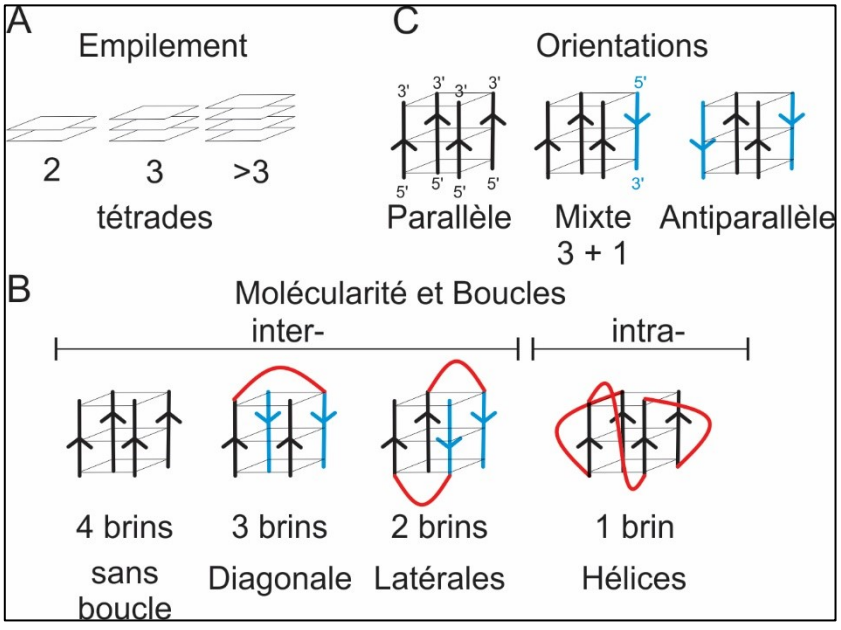



**Figure 5** – Diversité de molécularité, d'orientation et de boucles des G4.

(*Légende Figure 5*) (A) Les G4 peuvent être formés de l'empilement de 2, 3 ou plus tétrades. (B) Les séries de G essentielles à la formation des tétrades et dont le squelette forme les « arêtes » de la structure G4 peuvent être fournies par plusieurs brins distincts. Cela résulte en des molécularités différentes. Les boucles indiquées en rouge, qui relient les séries de G, peuvent être de différents types. (C) Le sens 5' vers 3' du squelette phosphodiester reliant les séries de G détermine si elles sont orientées de façons parallèle, antiparallèle ou mixte.

### G-quadruplex d'ARN

Les structures G4 sont donc très diversifiées, avec leur empilement, topologie, orientation et type des boucles différents. Cependant, cette diversité est plus limitée en ce qui concerne les G4 formés d'ARN, les **rG4**. En effet, la présence du groupement hydroxyle (2'-OH) sur le ribose entraîne des contraintes d'orientation entre le sucre et la base azotée. Contrairement à l'ADN où le lien glycosidique est plus flexible, le 2'-OH du ribose entraîne la conformation C3'-endo du sucre et force l'adoption d'un lien glycosidique *anti* (**Figure 6**).

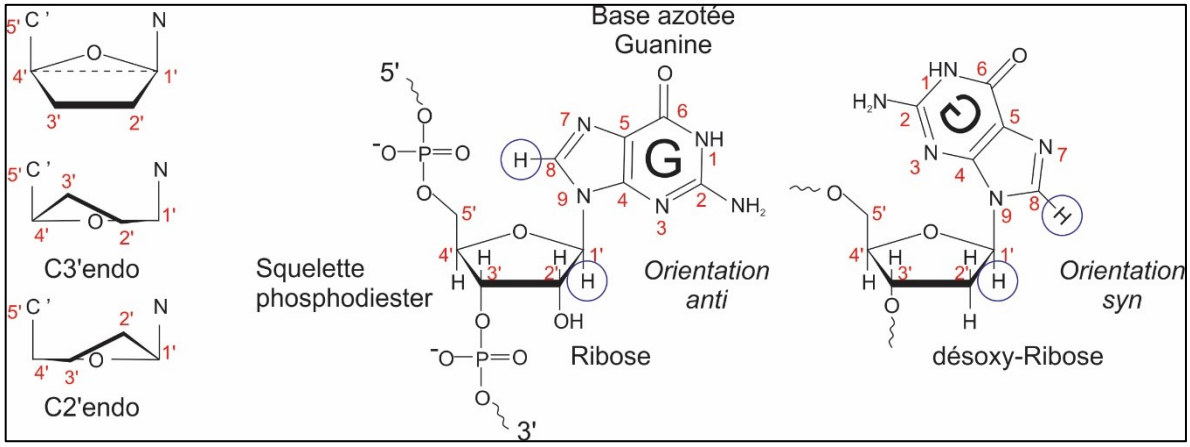



**Figure 6** – Conformations du ribose et orientations du lien glycosidique.

À gauche de la figure sont représentées les conformations C3'endo et C2'endo du ribose et à droite les orientations *anti* et *syn* de la guanine par rapport au ribose.

Il résulte de ces deux contraintes réunies que les G4 d'ARN intramoléculaires adoptent presque exclusivement une formation parallèle avec des boucles de type « hélice » tel que montré à la **Figure 4C** (Halder et Hartig, 2011). Cette structure qui a une apparence semblable à un disque (*Disc-like shape*) est très compacte, avec des tétrades très rapprochées les unes des autres, permettant des boucles pouvant être composées d'un seul nucléotide (Hazel *et al.*, 2004). La seule exception connue d'un G4 d'ARN antiparallèle est l'élucidation de la structure *in vitro* par résonance magnétique (RMN) de la séquence d'ARN télomérique TERRA (*telomeric repeat-containing RNA*) modifiée ou non par des 8-bromoguanosines

(Xiao *et al.*, 2017, 2018). Pour le moment, il n'y a pas d'évidence de la formation de ce type de topologie *in cellulo*. La topologie parallèle des rG4 reste considérée comme dominante.

**Stabilité du G4**

La structure G4 est une structure extrêmement stable. Cette stabilité est obtenue à plusieurs niveaux. En premier lieu, la stabilité résulte de la formation des 8 ponts hydrogène par tétrades, nombre qui est multiplié par le nombre de tétrades. En second lieu, une stabilisation électrostatique supplémentaire est obtenue par la présence du cation monovalent au centre de la tétrade. Finalement, l'empilement π (*π-stacking*) vient ajouter une stabilisation globale supplémentaire. Il en découle qu'en conditions physiologiques, soit en présence de 100 mM d'ions potassium $K^+$, la température de dénaturation ($T_m$) des G-quadruplex est très élevée pouvant aller de 70° à 95 °C et même plus (Gomez *et al.*, 2010).

Une réaction se déroule spontanément lorsque l'énergie libre des produits est inférieure à celle des réactifs, l'équilibre des systèmes tendant vers une stabilisation. En termes thermodynamiques, une réaction spontanée possède une différence d'énergie libre de Gibbs (ΔG) négative. Ce postulat est décrit par la formule $\Delta G=\Delta H-T\Delta S$, ou H signifie l'enthalpie (le nombre de liaisons), T la température et S, l'entropie (le désordre). Donc, en solution aqueuse, à température physiologique, la réaction de formation d'un G4 est spontanée. Elle consiste en la transition du réactif initial, le brin d'ARN non replié instable, au produit final, le repliement G4 très stable. La réaction est favorisée enthalpiquement par l'augmentation des diverses liaisons du G4 mentionnées précédemment (ponts hydrogène, l'empilement π, interactions électrostatiques). L'augmentation de l'ordre résultant de la structure entraîne une pénalité entropique (ΔS positif) compensée partiellement par la dé-solvatation des cations qui vont se coordonner entre les tétrades (Zaccaria et Fonseca Guerra, 2018).

D'un point de vue thermodynamique, la formation des G4 est donc spontanée et ceux-ci sont plus stables que les structures secondaires canoniques. Par contre, d'un point de vue cinétique, les G4 se forment moins rapidement en solution. Les G4 intramoléculaires canoniques d'ADN et d'ARN, formés de l'empilement de 3 tétrades, se replient dans l'ordre de 60 millisecondes (ms) en concentration favorable de $K^+$. Ce processus comporte deux intermédiaires : la formation d'une épingle à cheveux (2 séries de G qui interagissent), puis d'un triplex pour obtenir finalement le G4 replié. La formation de G4 formé de 2 tétrades,

d'une durée d'environ 700 ms, est plus lente que le G4 à 3 tétrades. La formation peut dépasser 100 s pour un G4 doublet lorsque les boucles sont plus longues (Zhang et Balasubramanian, 2012). Les deux premiers temps de formation mentionnés sont compatibles avec les vitesses d'élongations des polymérases et confirment que le repliement G4 peut s'effectuer durant la réplication et la transcription et être conservé durant la durée de vie de l'espèce ARN transcrite. Afin de comparaison, une structure tige-boucle d'ARN de 21 nt se replie en 0,1 ms seulement (Zhang et Chen, 2002). Donc, bien que beaucoup moins stable, une structure Watson-Crick alternative impliquant les G du G4 peut être favorisée à l'équilibre puisqu'elle sera adoptée plus rapidement par le brin d'ARN.

*Différences entre G4 d'ADN et d'ARN*

Un G4 formé d'une séquence d'ARN est plus stable que celui adopté par une séquence équivalente en ADN (Mergny *et al.*, 2005 ; Zaccaria et Fonseca Guerra, 2018). Cela est dû à plusieurs facteurs. D'abord la présence du ribose et de son 2'-OH permet la formation de liens hydrogène stabilisateurs supplémentaires avec le squelette phosphodiester. Cela permet aussi un meilleur empilement π des tétrades (Olsen et Marky, 2009). De plus, tel que mentionné précédemment, l'absence de diversité de topologie pour les G4 d'ARN augmente leur stabilité, puisqu'il n'y a pas d'interconversion possible entre les topologies selon les diverses conditions en solution comme c'est le cas pour les G4 ADN. La topologie parallèle avec des boucles de type hélice est la plus stable de toutes (Joachimi *et al.*, 2009). Un autre facteur important est la molécularité en cellules de l'ARN par rapport à l'ADN. L'ARN est biologiquement présent sous forme simple-brin et peut adopter diverses structures intramoléculaires alors que l'ADN est toujours en présence de son brin complémentaire. Donc, la formation de G4 d'ADN est limitée par la constante compétition avec la formation de la structure double-brin canonique.

**Facteurs influençant la formation et la stabilité des G-quadruplexes**

La formation de toute structure secondaire, canonique ou non, peut être influencée par de multiples facteurs. Ceux-ci peuvent être des facteurs agissant en *cis*, c'est-à-dire qu'ils peuvent provenir de la composition de la molécule d'ARN elle-même. Les facteurs peuvent aussi être extérieurs, retrouvés dans l'environnement qui entourent la molécule d'ARN et qui

peuvent interagir avec elle. Dans ce cas, ils sont appelés facteurs *trans*. Les facteurs principaux affectant la formation et la stabilité des G4 sont résumés à la **Figure 7**.

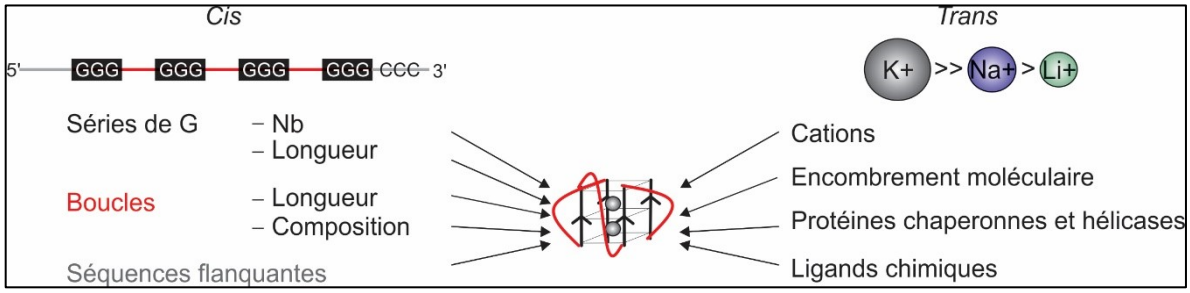



**Figure 7** – Facteurs *cis* et *trans* influençant le repliement rG4.

Les facteurs *cis* sont les éléments de la séquence intrinsèque du rG4 ainsi que des séquences adjacentes. Le motif G4 est représenté à gauche sous une forme linéaire avec les séries de G représentées par les 4 rectangles noirs, les nucléotides des boucles sont représentés par les lignes rouges entre les séries de G et les séquences adjacentes sont représentées par les lignes grises en 5' et en 3' avec une emphase sur les séries de C. Les facteurs *trans* sont les éléments extérieurs à la séquence qui peuvent interagir avec la structure rG4. Ils sont énumérés à droite et les cations monovalents permettant de stabiliser les tétrades sont représentés en ordre du plus stabilisateur au moins stabilisateur. La taille du cercle représente la taille du rayon ionique des cations.

*Séquence et motif G4*

À partir des fondements chimiques de la formation de la structure G4, il est possible d'identifier les facteurs prérequis essentiels à la formation d'un G4 intramoléculaire. Tout d'abord, la séquence doit posséder 4 répétitions d'une série de deux G consécutifs ou plus, séparées par des boucles de n'importe quel nucléotide (N = A, T, C, G, U).

*Empilement des tétrades de G*

Tel que mentionné précédemment, la stabilité des G4 est fortement influencée et proportionnelle au nombre d'empilements de tétrades. Un G4 avec des triplets de G (3 tétrades empilées) sera plus stable qu'un G4 formé de doublets de G (Pandey *et al.*, 2013). Bien que chimiquement parlant, l'empilement de tétrades n'ait pas de limite, biologiquement, des structures d'une telle stabilité sont à éviter, car l'équilibre entre le repliement et la dénaturation serait difficile à obtenir. À ce jour, le nombre maximal de tétrades empilées observé pour des G4 d'ADN et d'ARN de séquence naturelle avec des effets biologiques connus est de quatre. Il s'agit entre autres de la répétition nucléotidique GGGGCC retrouvée dans le gène C9ORF72 (Conlon *et al.*, 2016 ; Zhou *et al.*, 2015). Il existe tout de même plusieurs régions du génome et du transcriptome possédant 4 répétitions de séries de 4 G

consécutifs ou plus dont le potentiel G4 n'a pas été encore confirmé ou infirmé expérimentalement (Huppert et Balasubramanian, 2005 ; Beaudoin et Perreault, 2010).

En plus du nombre de tétrades empilées, donc de la longueur des séries de G, il faut aussi considérer le nombre de répétitions de séries de G. Il en faut un minimum de quatre, mais plusieurs séquences avec le potentiel d'adopter un G4 (**potentiel G4, PG4**) en comportent 5, 6 ou plus situées à proximité (Burge *et al.*, 2006). Après les quatre essentielles, chaque répétition supplémentaire d'une série de G consécutifs entraîne une augmentation des combinaisons différentes possibles permettant la formation d'un G4. Cela augmente la diversité de l'ensemble de G4 présents pour une même séquence. Ces diverses combinaisons de 4 séries peuvent donner des G4 ayant des stabilités différentes dues aux boucles différentes qui en résultent.

*Taille et composition des boucles*

La taille, la composition et l'organisation des boucles reliant les séries de G ont des effets sur la stabilité de la structure. Des analyses systématiques des boucles possibles ont démontré que les boucles courtes donnent des G4 plus stables (Hazel *et al.*, 2004 ; Risitano et Fox, 2004 ; Zhang *et al.*, 2011a ; Bugaut et Balasubramanian, 2008). De plus, les G4 sont plus stables lorsque les 3 boucles sont de la même longueur ou si seulement la boucle centrale est plus longue (Guédin *et al.*, 2009; Rachwal *et al.*, 2007b). Dans ces études, les boucles sont de longueur restreinte, souvent de 1 à 3 nt, les boucles supérieures à 3 nt sont considérées comme « longues » et ne dépassent rarement 7 nt. Pour les G4 d'ADN, la longueur des boucles peut affecter la topologie adoptée par le G4. Des boucles plus courtes favorisent la conformation parallèle (Tippana *et al.*, 2014). Cet aspect ne s'applique pas aux rG4 puisque ceux-ci adoptent uniquement la topologie parallèle due aux raisons mentionnées préalablement.

Il est possible de mesurer l'effet de la composition nucléotidique des boucles sur la stabilité en comparant des séquences PG4 avec des tailles de boucles constantes. Les G4 parallèles avec des boucles composées d'un seul nucléotide adénine sont moins stables que ceux avec des boucles formées d'une seule pyrimidine, thymine ou cytosine (Rachwal *et al.*, 2007a ; Guédin *et al.*, 2008). Des G4 avec des boucles d'un seul nucléotide A, C ou T sont aussi plus stables qu'une boucle formée d'un G. La substitution d'une boucle T par un U résulte aussi en un G4 plus stable, ce qui s'ajoute aux autres preuves d'une plus grande

stabilité des G4 d'ARN que d'ADN (Olsen *et al.*, 2009). La stabilité du G4 peut aussi être affectée en permutant la position d'une même séquence entre la première, deuxième ou troisième boucle du G4 (Cheng *et al.*, 2018). Cela s'explique par les différentes possibilités d'empilement et d'interactions électrostatiques que les nucléotides des boucles peuvent former avec le cœur tétrade du G4 selon leur position. L'impact de la composition d'une boucle est aussi important pour des rG4. La modification de la boucle centrale de 6 nt du rG4 retrouvés dans l'ARNm PITX1 par une séquence aléatoire de même longueur n'affecte pas le repliement rG4, mais affecte sa vitesse de migration dans un gel natif, ainsi que sa liaison à un composé chimique et à une hélicase spécifiques aux rG4 (Ariyo *et al.*, 2017).

Il est important de noter que la majorité des études visant à isoler spécifiquement l'effet de chaque caractéristique principale des G4 sur la stabilité de la structure ont été effectuées en utilisant des séquences G4 d'ADN plutôt que d'ARN. Plusieurs conclusions peuvent être valables pour les deux types d'acides nucléiques, mais certaines différences ont pu être occultées par le manque d'études systématiques avec des séquences ARN.

*Cations monovalents*

L'orientation vers le centre des groupements carboxyl de chacune des guanines résulte en une charge partielle négative au centre de la tétrade. Pour stabiliser la tétrade, cette charge négative doit être compensée par une charge positive, soit un cation. Divers ions de métaux lourds, par exemple le rubidium ($Rb^{+}$) et le strontium ($Sr^{3+}$), permettent de stabiliser les tétrades de G en se coordonnant dans le canal central (Neidle, 2012). Les ions ammonium $NH_4^{+}$ aussi sont capables de compenser les charges négatives des tétrades. Les cations les plus efficaces pour la stabilisation de tétrades sont cependant les cations monovalents dont les concentrations sont les plus abondantes en cellules : le potassium ($K^{+}$) et le sodium ($Na^{+}$) avec des concentrations intracellulaires respectives de 140 mM et 10 mM (Meyers, 2004). La stabilisation est différente selon la taille du rayon ionique du cation. Le $K^{+}$ a un rayon trop large (1,33 Å) pour être égal au plan de la tétrade, c'est pour cela qu'il se coordonne entre deux plans. Tandis que le $Na^{+}$ avec un rayon un peu plus petit (0,97 Å) peut se retrouver entre deux plans ou vis-à-vis un plan. Un autre ion de métal alcalin, le lithium ($Li^{+}$) a quant à lui un rayon ionique trop petit (0,68 Å) pour stabiliser suffisamment la tétrade (Fay *et al.*, 2017 ; Hardin *et al.*, 1991, 1992 ; Shannon, 1976).

En général, les G4 d'ADN sont plus sensibles à la présence du cation $Na^+$, alors que les rG4 sont stabilisés principalement par le $K^+$. Selon la présence d'un ou l'autre de ces deux cations monovalents, le G4 issu de la séquence des répétitions télomériques (TTAGGG) adoptera une conformation antiparallèle ou parallèle. Les rG4 à cause du groupement 2'-OH du ribose, adopte uniquement la formation parallèle et donc l'effet de la nature du cation présent n'entraîne pas de conséquence topologique (Lane *et al.*, 2008 ; Halder et Hartig, 2011). Les structures secondaires canoniques sont généralement stabilisées par la présence de magnésium ($Mg^{2+}$). Les G4 et rG4 sont les seules structures secondaires dépendantes du $K^+$ pour leur stabilité.

*Encombrement moléculaire*

En condition physiologique, les acides nucléiques ne sont pas « dilués » en solution. L'encombrement élevé est dû à la présence abondante de plusieurs macromolécules, d'osmolytes et de petites molécules organiques dans le cytoplasme. Les structures secondaires et tertiaires de l'ARN sont stabilisées par un tel environnement encombré (Nakano *et al.*, 2014). C'est le cas aussi pour les structures G4. Donc, naturellement la formation de G4 en cellule est favorisée. En condition *in vitro,* on peut simuler un environnement moléculaire encombré en ajoutant des agents d'encombrement tel que le polyéthylène glycol (PEG) ou des polysaccharides comme le dextrane. C'est ainsi que l'effet de l'encombrement sur le repliement G4 a été étudié. Il a été constaté qu'en combinaison avec la séquence primaire et à la présence de cations, la topologie adoptée par les G4 d'ADN est aussi affectée par les conditions d'encombrement. En général, l'augmentation de l'encombrement stimule la conformation parallèle (Miyoshi *et al.*, 2013). De plus, l'augmentation de l'encombrement moléculaire favorise la dissociation de duplex en faveur de la formation de G4 (Kumar et Maiti, 2005 ; Miyoshi *et al.*, 2004). Puisque les rG4 n'adoptent que la conformation parallèle et qu'ils ne sont pas aussi polymorphiques que les G4 d'ADN, l'encombrement moléculaire n'affecte donc pas leur conformation. Par contre, ils sont eux aussi favorisés par rapport à une structure duplex, et stabilisés par l'augmentation de l'encombrement moléculaire (Zhang *et al.*, 2010a).

*Compétition avec des structures secondaires canoniques*

Tel que mentionné précédemment, la cinétique de formation de structure secondaire Watson-Crick est beaucoup plus rapide que celle de la formation de G4. La présence d'une séquence de nucléotides complémentaires au motif PG4 constitue donc un facteur compétitif nuisible à la formation de G4. Si la structure Watson-Crick alternative est en elle-même assez stable, une étape limitante dans la conversion de la structure secondaire duplex vers le G4 sera la dénaturation de cette tige-boucle (Kuo *et al.*, 2015b). Donc, malgré le fait que les G4 soient beaucoup plus stables d'un point de vue thermodynamique, si une conformation Watson-Crick de stabilité intermédiaire est obtenue pour une séquence, celle-ci sera formée plus tôt et constituera un puits énergétique intermédiaire empêchant la formation du G4 puisque l'énergie de transition nécessaire entre les 2 états sera trop élevée. Il est donc logique que la présence d'une séquence complémentaire au G4, tel le brin complémentaire dans l'ADN, ou des séquences adjacentes riches en cytosines sur un brin d'ARN viennent s'apparier avec les guanines et nuire au repliement intramoléculaire du G4 ou du rG4. En effet, la propension à adopter le repliement G4 diminue plus on augmente la longueur des séquences adjacentes en 5' et en 3' du motif PG4, car la possibilité de structures secondaires alternatives augmente (Saxena *et al.*, 2010). De plus, la stabilité mesurée par dénaturation thermique d'un motif G4 diminue lorsqu'on augmente son contexte nucléotide, car plusieurs structures alternatives moins stables sont alors présentes dans l'ensemble (Arora *et al.*, 2009).

Les facteurs influençant la formation de G4 initialement mentionnés tels que la taille des boucles ainsi que la présence de cations peuvent influencer l'interconversion entre duplex et quadruplex. Des boucles plus longues entraînent une plus grande possibilité de compétition avec une structure duplex complémentaire (Kumar *et al.*, 2008). La transition d'une conformation tige-boucle à G4 est aussi favorisée par la concentration de différents cations en solution. Lorsqu'une séquence d'ARN peut adopter de façon mutuellement exclusive un G4 ou une tige-boucle, la présence du cation divalent $Mg^{2+}$ favorise la tige-boucle alors que le cation monovalent $K^{+}$ favorise le rG4 (Bugaut *et al.*, 2012). Ces deux cations sont les plus abondants dans les cellules. Dans les conditions physiologiques cellulaires, on assume donc la présence d'un ensemble de structures secondaires possibles pour une même séquence PG4, incluant G4 et duplex, dont l'équilibre peut changer de façon dynamique selon les différentes conditions. L'équilibre entre les conformations alternatives peut être déplacé dans un sens

comme dans un autre selon la présence de protéines et des hélicases liant les séquences d'acides nucléiques, ainsi qu'en utilisant des composés synthétiques, ou en fournissant de façon ectopique des séquences d'oligonucléotides complémentaires.

*Liaison de petites molécules et oligonucléotides complémentaires*

Avec leur topologie unique et la grande surface aromatique d'interaction que fournissent les tétrades des G4, il est possible de cibler spécifiquement cette structure secondaire grâce à de petites molécules chimiques. Celles-ci peuvent s'empiler sur la tétrade du dessus et stabiliser grandement la structure secondaire. Un pan complet du domaine d'étude des G4 est consacré au développement de ce type de composés pour détecter ou moduler les G4. Ceux-ci seront décrits plus en détail dans les sections détection et ciblage des G4. On peut cibler les G4 avec des petites molécules artificielles, mais certaines biomolécules naturellement présentes dans les cellules peuvent aussi avoir une affinité de liaison avec les tétrades. C'est le cas des porphyrines telles que l'hémine (Kosman et Juskowiak, 2016).

Une façon encore plus intuitive d'influencer le repliement d'une structure secondaire est d'utiliser une séquence d'acide nucléique complémentaire. C'est la stratégie derrière l'utilisation d'oligonucléotides antisens (*antisens oligonucleotide,* ASO). Ce sont souvent des séquences longues de 15 à 20 nt, modifiées chimiquement afin d'éviter leur dégradation par les nucléases. Cela fonctionne tel que décrit pour la compétition avec des séquences complémentaires adjacentes. La formation du duplex entre la séquence PG4 et l'ASO préviendra la formation du G4. L'avantage de l'ASO est qu'il peut être très spécifique pour un seul G4 grâce à la complémentarité de séquence (Rouleau *et al.*, 2015). À l'inverse, il est aussi possible de favoriser la formation de G4 en utilisant des courtes séquences d'oligonucléotides. Une première façon est d'utiliser des ASO qui viennent séquestrer des séquences adjacentes compétitrices, comme des séries de cytosines. Une seconde façon est de stimuler la formation de rG4 intermoléculaire dans des régions où le nombre de séries de G est insuffisant (<4) en fournissant des séries de G supplémentaires sous forme d'oligonucléotides courts. Cela permettra par exemple la formation d'un G4 intermoléculaire entre un ARNm qui possède 2 séries de 3 G consécutifs et l'oligonucléotide qui fournira les 2 autres séries de G nécessaires (Ito *et al.*, 2011). Il existe même des combinaisons entre courts oligonucléotides et ligands chimiques spécifiques aux G4. Par exemple, un oligonucléotide formé d'une série de G conjugué à un ligand permettra de restaurer la

formation d'un G4 dans une séquence du génome où une série de G du G4 original serait mutée. La séquence de l'oligonucléotide apporterait la spécificité et compléterait la série de G et le ligand apporterait une stabilisation supplémentaire de la structure (Takahashi *et al.*, 2018).

*Protéines et hélicases*

Les protéines et les hélicases sont des facteurs de régulation en *trans* des structures secondaires qui sont présents naturellement dans les conditions biologiques. Les G4 sont des structures secondaires si stables qu'elles ne peuvent être dépliées facilement par les polymérases ou les ribosomes, et qui nécessitent donc la présence de facteurs protéiques pour favoriser leur dénaturation. Les hélicases sont les protéines responsables de déplier les structures secondaires d'acides nucléiques. La majorité des hélicases utilisent l'hydrolyse de l'ATP comme substrat énergétique afin de se déplacer tout au long d'une séquence repliée et ainsi défaire les structures secondaires présentes. Il existe des hélicases spécifiques pour l'ADN ou l'ARN et qui ont des affinités particulières pour certains motifs et types de structure, alors que d'autres sont non spécifiques et résoudront n'importe quelle structure secondaire que ce soit des duplex, triplex, jonctions ou quadruplex. À ce jour, une dizaine de protéines hélicases ont été identifiées avec une affinité pour les structures G4 (Mendoza *et al.*, 2016). Les plus étudiées sont les hélicasesPif1, RecQ, FANCJ, BLM (*Bloom syndrome protein*) et WRN (*Werner syndrome protein*) qui régulent la formation et la dénaturation des structures G4 d'ADN principalement durant la réplication du génome et la transcription. À ce jour, trois hélicases sont caractérisées avec une activité spécifique pour les rG4. Ce sont les hélicases DHX36, DHX9 et DDX21. L'hélicase de rG4 la mieux connue est DHX36, aussi appelée RHAU (*RNA helicase binding AU-rich element*), MLEL1 (*MLE-like protein* 1), ou G4R1 (*G4 resolvase 1)*. C'est une hélicase ADN et ARN de la famille des *DEAH-Box* avec une activité résolvase de sens 3' vers 5'. DHX36 est l'hélicase avec activité de résolution de G4 prédominante en cellule (Creacy *et al.*, 2008). Elle reconnaît des rG4 situés autant dans les régions non traduites (UTR) que dans les régions codantes des transcrits ainsi que dans des ARN non codants. DHX36 a des rôles importants dans le développement du cœur, l'hématopoïèse et l'embryogenèse (Chen *et al.*, 2018a).

La protéine eIF4A est l'hélicase présente dans le complexe d'initiation de la traduction. Une étude a démontré que dans des cellules cancéreuses de leucémie aiguë

lymphoblastique à leucocyte T (T-ALL, *T-Cell acute lymphoblastic leukaemia*), les transcrits les plus dépendants de eIF4A pour leur traduction possédaient un motif PG4 suggérant une affinité de cette hélicase pour les rG4 (Wolfe *et al.*, 2014). Les autres hélicases connues avec des activités de résolutions de rG4 sont DHX9 et DDX21. Plusieurs de ces hélicases sont surexprimées ou dérégulées dans certains cancers (Fuller-Pace, 2013), ce qui suggère un effet possible du dérèglement de l'équilibre entre G4 repliés ou non dans la carcinogenèse.

D'autres protéines liant l'ARN (les **RBP**, ***RNA-binding proteins***) sans activité hélicase, peuvent reconnaître et stabiliser la formation des rG4. Une des mieux connues est FMRP (*Fragile X mental retardation protein*). Cette RBP régule la localisation et la traduction des transcrits dans les neurones et reconnaît certaines de ces cibles grâce à la présence de rG4 dans leur 3'UTR. Plusieurs autres RBP décrites comme liant les rG4 interagissent plutôt avec les séquences riches en G qui sont enclines à la formation de G4 pour justement prévenir leur formation. Le scénario inverse est aussi possible, soit la présence de protéines qui lient les régions poly(C) ce qui favoriserait le repliement G4 en inhibant le repliement de structures secondaires compétitrices (Beaudoin et Perreault, 2010). Le **Tableau 1** présente un résumé des protéines et des hélicases impliquées dans la régulation et les processus biologiques impliquant des structures rG4.

**Tableau 1** Protéines liant les rG4

| Fonction | Protéine | Mécanisme relié aux rG4 |
|---|---|---|
| **Hélicase** | DHX36,(RHAU, MLEL1, G4R1) | Biologie des télomères, repliement de l'ARN de la télomérase (Booy *et al.*, 2012) |
| | | Régulation de la traduction (Chen *et al.*, 2018a ; Murat *et al.*, 2018) |
| | DDX21 | Régulation de l'expression protéique (McRae *et al.*, 2017) |
| | DHX9 (RHA, NDH II) | Transcription (résoudre R-loop et G4 dans les transcrits)(Chakraborty et Grosse, 2011) |
| | eIF4A | Régulation de la traduction (Wolfe *et al.*, 2014) |
| | MOV10 | Régulation de la traduction médiée par miARN (Kenny *et al.*, 2014) |
| **RNA-binding protein, RBP** | Nucleolin | Stabilité des ARNm et Traduction (von Hacht *et al.*, 2014) |
| | FMRP | Stabilité des ARNm, Localisation et Traduction (Vasilyev *et al.*, 2015) |
| | Aven | Traduction (Thandapani *et al.*, 2015) |
| | hnRNP A2 | Régulation de la traduction (Khateb *et al.*, 2007) |
| | CNBP/ZNF9 | Régulation de la traduction (Benhalevy *et al.*, 2017) |
| | Grsf1 | Régulation de la traduction suggérée (Nieradka *et al.*, 2014) |
| | YB1 | Traduction, via fragments de tRNA (tRNA-derived stress induced fragments)(Ivanov *et al.*, 2014) |
| | hnRNP A1 | Épissage et Traduction (Cammas *et al.*, 2016 ; Zamiri *et al.*, 2014) |
| | hnRNP A3 | Épissage (Conlon *et al.*, 2016) |
| | hnRNP H | Épissage, répétitions GGGGCC associées à ALS (Conlon *et al.*, 2016) |
| | hnRNP F | Épissage (Huang *et al.*, 2017) |
| | FMR2 (AFF2) | Épissage (Bensaid *et al.*, 2009) |
| | AFF3, AFF4 | Épissage (Melko *et al.*, 2011) |
| | U2AF65 | Épissage suggéré (pas d'essai fonctionnel)(von Hacht *et al.*, 2014) |
| | SRSF1 (ASF/SF2) | Épissage suggéré (pas d'essai fonctionnel)(von Hacht *et al.*, 2014) |
| | FUS/TLS | Biologie des télomères, lie TERRA (Takahama et Oyoshi, 2013) |
| | TRF2 | Biologie des télomères, lie TERRA (Biffi *et al.*, 2012) |
| | Lin28 | Stabilité des ARNm et des miARN(O'Day *et al.*, 2015) |
| | DDX3X<br>DDX5<br>DDX17<br>GRSF1<br>NSUN5 | Régulation post-transcriptionnelle suggérée, partenaires d'interactions de rG4 en 5'UTR récemment identifiés (Herdy *et al.*, 2018a) |

Références supplémentaires : (Mendoza *et al.*, 2016 ; Fay *et al.*, 2017 ; Sauer et Paeschke, 2017)

Comme les rG4 ne sont pas uniformes, que chacun peut posséder des particularités structurales selon son nombre de tétrades, la taille et la composition des boucles, etc., les hélicases et les RBP identifiées ne reconnaissent pas nécessairement tous les rG4 de manière équivalente, mais ont plutôt des affinités pour certaines sous-catégories de rG4. Par exemple, la nucleolin a plus d'affinité pour des G4 avec des boucles plus longues (Lago *et al.*, 2017) tandis que les motifs PG4 des transcrits dont la traduction est dépendante de eIF4A ont des séries de 2 G (Wolfe *et al.*, 2014). La RBP NSUN5 a une affinité beaucoup plus grande pour le rG4 du transcrit NRAS que pour celui du transcrit BCL-2, bien que les deux rG4 aient des motifs semblables, mais des boucles de composition différentes (Herdy *et al.*, 2018b). La plupart des protéines liant les rG4 ont été identifiées grâce à des techniques de *pull-down*, en utilisant comme appât des séquences rG4 bien caractérisées *in vitro*, souvent associées à des transcrits importants dans la progression tumorale (Serikawa *et al.*, 2017). Ces protéines reconnaissent donc des rG4 « canoniques » et on ne sait pas si elles peuvent reconnaître les rG4 atypiques qui seront décrits plus loin dans l'introduction. La régulation du repliement rG4 en *trans* par des protéines est donc un aspect biologique très important pour moduler leurs fonctions biologiques, et les rôles de ces facteurs protéiques dans les processus biologiques impliquant les rG4 seront décrits plus en détail dans la section sur les rôles biologiques des rG4.

En somme, la formation des rG4 est dépendante à la fois de leur séquence intrinsèque, de leurs séquences adjacentes et des facteurs biologiques (cations, encombrement moléculaire, protéines) ou artificiels (ligands chimiques, ASO) présents dans la cellule à un moment précis. L'interaction entre tous ces aspects résulte en la formation d'un équilibre dynamique entre les différentes conformations de structures rG4 possibles ou de ses structures alternatives canoniques.

**Prédiction des G4**

L'intérêt scientifique envers la structure G-quadruplex a connu une hausse à partir des années 2000, dès le moment où les caractéristiques fondamentales de la structure, ainsi que les rôles des G4 d'ADN dans l'inhibition de la télomérase et de la transcription de proto-oncogènes ont été mieux connus. De plus, les possibilités de cibler cette structure

pharmacologiquement ont entraîné une recherche effrénée afin de découvrir l'ensemble des G4 présents dans le génome.

**Recherche de motifs G4 canoniques**

La première méthode de prédiction utilisée a été l'analyse de séquences afin d'identifier un motif G-quadruplex. Cette façon reste à ce jour la méthode la plus fortement utilisée. Elle a été développée et éprouvée simultanément par deux groupes de recherche (Huppert et Balasubramanian, 2005 ; Todd *et al.*, 2005). Basé fondamentalement sur les caractéristiques de la séquence G4 la plus étudiée, celle des répétitions télomériques, et sur l'étude de la stabilité *in vitro* de plusieurs G4 d'ADN artificiels, le motif G-quadruplex suivant a été défini :

$$\mathbf{G}_x\text{-}N_y\text{-}\mathbf{G}_x\text{-}N_y\text{-}\mathbf{G}_x\text{-}N_y\text{-}\mathbf{G}_x, \text{ où } x= 3 \text{ à } 5,\ y=1 \text{ à } 7$$

Ce motif, qu'on appelle le motif G4 consensus ou canonique, souvent intitulé *PQS* pour *probable quadruplex sequence* ou PG4 dans la littérature, représente les éléments essentiels d'un G4 intramoléculaire, soit minimalement 4 répétitions de 3 à 5 Gs (les séries de G) qui formeront les tétrades empilées, reliées par 3 boucles formées de n'importe quel type de nucléotide (A, T, U, G, ou C) d'une longueur minimale de 1 nt et allant jusqu'à 7 nt. Les G4 avec des boucles dépassant la longueur de 7 nt ayant une stabilité beaucoup plus faible, cette limite a été imposée. En utilisant *Quadparser*, c'est-à-dire un algorithme qui interroge le génome afin d'identifier toutes les régions qui respectent ce motif canonique, plus de 376 000 G4 potentiels (PG4) ont été identifiés dans le génome humain. Ceux-ci se retrouvent de façon non aléatoire dans le génome, étant enrichis dans les promoteurs, ainsi que dans les régions 5' et 3' UTR entre autres (Huppert et Balasubramanian, 2005).

Cette méthode de prédiction ou d'identification de PG4 est simple et intuitive, cependant elle reste très stricte en limitant grandement le nombre de G dans les séries et la taille des boucles, la taille de la région PG4 étant de minimum 15 nt à maximum 41 nt. Plusieurs autres variations sur ce motif ont donc été proposées pour augmenter la diversité des prédictions PG4, et des outils ont été développés pour faciliter la recherche. Entre autre, l'outil *QGRSmapper,* pour *Quadruplex-forming G-rich sequences* (Kikin *et al.*, 2006), permet à l'utilisateur de varier les paramètres du motif, soit le nombre minimal ou maximal de G dans les séries (2 à 6), la taille des boucles (1 à 36), ou la taille maximale du motif

complet de 10 à 45 nt. Selon les paramètres les plus stricts aux moins stricts de cet outil, de 197 000 jusqu'à 2 391 000 régions PG4 sont identifiées dans le génome humain. La force de cet outil est l'ajout d'un système de score permettant de juger de la probabilité du motif retrouvé de former réellement ou non un G4, appelé le G-score. Ce score tient compte de trois facteurs : premièrement, le nombre de G dans les séries, un nombre plus élevé étant associé à un G4 plus stable ; deuxièmement, la taille des boucles, des boucles plus courtes étant plus stables que des boucles plus longues ; et, troisièmement, la symétrie des boucles, trois boucles de tailles identiques étant jugées plus stables que trois boucles de tailles variables. Le score s'échelonne de 0 à 105. Bien qu'informateur, ce score n'a pas été évalué empiriquement : il n'a pas été mesuré si un score supérieur représentait bel et bien une augmentation de la stabilité structurale en termes d'énergie libre minimale de repliement (Mfe*, minimum free energy*). De nombreux autres outils de prédictions de G4 reposant tous sur la recherche du motif canonique *PQS*, ou sa version « étendue » avec des boucles maximales allant jusqu'à 12 nt plutôt que 7, ont été développés et ceux-ci sont résumés dans la catégorie « Recherche de motif canonique » dans le **Tableau A1** à l'**Annexe 1**.

**Stabilité de la structure secondaire**

D'autres outils ont été créés justement dans le but d'évaluer et de prédire la stabilité de la structure G4. Cela étant toujours fondé sur l'idée qu'un G4 stable a plus de probabilité de se former. Le premier outil de ce genre, *Quadpredict,* a été entraîné avec une banque de données de G4 ADN courts dont la stabilité a été déterminée expérimentalement *in vitro* (Stegle *et al.*, 2009). Cet outil ne permet pas de prédire si une séquence forme ou non un G4, mais permet plutôt de comparer les différents degrés de stabilités des PG4 d'ADN.

Par contre, basé sur le principe de la stabilité, le domaine complet de la prédiction de structure secondaire d'ARN est fondé sur un modèle thermodynamique où une structure avec des appariements plus stables a plus de probabilité à se replier qu'une structure alternative pouvant être formée par la même séquence, mais de stabilité moindre. De nombreux outils existent afin d'estimer la stabilité et le repliement de structures secondaires d'ARN, tels que *RNAfold*, *RNAstructure* et *mFold* (Lorenz *et al.*, 2011 ; Reuter et Mathews, 2010 ; Zuker *et al.*, 1999). Un seul outil de prédiction de structure secondaire d'ARN inclut la prédiction de structure rG4. Celui-ci est l'outil *RNAfold*. L'option de prédiction de rG4 est offerte dans les paramètres avancés de l'outil. La méthode est basée sur la combinaison de la recherche d'un

motif G4 (*PQS*) avec un modèle thermodynamique simplifié (Lorenz *et al.*, 2013). D'abord, cet outil prédit la stabilité de l'ensemble des structures secondaires canoniques possibles d'une séquence d'ARN. Quand un motif PG4 est reconnu dans une séquence, une mesure de la stabilité de l'énergie libre minimum (Mfe) est calculée, basée sur le modèle thermodynamique des G4. Cette stabilité de structure avec rG4 est comparée avec l'ensemble de prédictions de structures canoniques. Le résultat de la prédiction sera la structure la plus stable entre ces différentes possibilités.

Les résultats de prédictions de structures secondaires peuvent être représentés de différentes façons : schématiquement, en représentation en arc (*arc-plot*) ou en notation *dot-and-bracket.* Dans la représentation en arc, la séquence orientées dans le sens 5' vers 3' est représentée par une ligne horizontale et les nucléotides qui sont appariés de façon canonique sont reliés par un arc. Cette notation ne peut s'appliquer pour les G4. Dans la notation *dot-and-bracket*, chaque nucléotide de la séquence est représenté par un point « . » ou une parenthèse ouverte ou fermée « ( », « ) ». Un point représente un résidu non apparié et les parenthèses les résidus qui sont appariés ensemble. Dans les cas des structures secondaires rG4, un nouveau symbole est ajouté dans l'outil *RNAfold* : le « + ». Il permet d'identifier les G faisant partie des tétrades prédites (**Figure 8**).

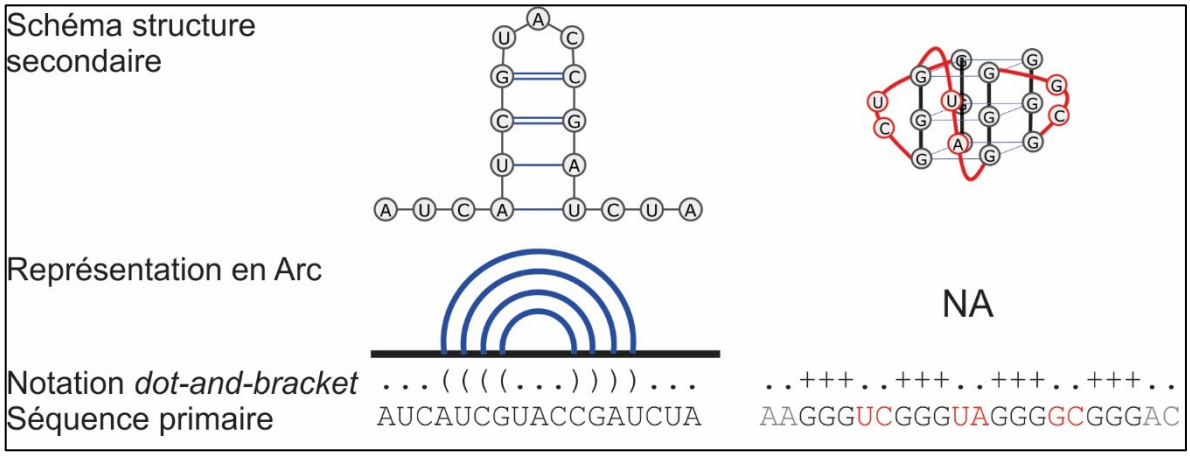



**Figure 8** – Deux types de représentations des structures secondaires d'ARN : en arc et en *dot-and-bracket.*

La représentation en Arc n'est pas applicable pour les rG4. Pour représenter les G impliqués dans des tétrades, la notation *dot-and-bracket* ajoute le symbole « + ».

**Densité des motifs PG4**

Une autre méthode beaucoup moins stricte et beaucoup plus intuitive a été utilisée à l'origine pour analyser la présence de motifs PG4 dans des régions régulatrices du génome. Afin d'identifier si les G4 étaient associés à certaines familles de gènes, Johanna Eddy et Nancy Maizels ont développé l'outil *G4P-calculator* (Eddy et Maizels, 2006). Cette méthode utilise une fenêtre défilante très large de 100 nt de long, avec un décalage de 20 nt entre chaque fenêtre, afin d'interroger le génome. La densité de PG4 est mesurée simplement comme le nombre de 4 séries ou plus contenant un minimum de 3 G consécutifs dans une fenêtre. Le tout résulte en une densité de PG4 en pourcentage, indépendant de la longueur, en ne tenant pas compte des boucles possibles, mais considère le contexte étant donné la largeur de la fenêtre. De façon surprenante, cette méthode obtient des résultats très similaires à l'outil *Quadparser* quant au nombre et à la localisation des régions riches en PG4. Elles ont pu identifier que la densité PG4 corrèle avec certaines classes fonctionnelles de gènes. Les PG4 sont enrichis dans les promoteurs de proto-oncogènes et déplétés dans les promoteurs de gènes suppresseurs de tumeurs. De plus, le premier intron en 5' des transcrits sont aussi des régions très denses en PG4 (Eddy et Maizels, 2008).

En général, on constate que la majorité des outils de prédictions ont été développés et validés en n'utilisant que des séquences d'ADN et des conditions *in vitro* plus ou moins physiologiques. Ceux-ci sont aussi strictement limités aux deux caractéristiques minimales des G4, les séries de G et les boucles.

**G-quadruplexes atypiques (non canoniques)**

Les méthodes de prédictions sont basées sur un motif de G4 canonique établi par les études de stabilité *in vitro* qui ont permis de délimiter la longueur des éléments essentiels de la séquence pouvant adopter un G4. Cependant, certaines preuves expérimentales démontrent que des G-quadruplexes « atypiques » ou divergents du motif consensus peuvent se former. D'abord, le nombre minimal requis de 3 G dans chacune des séries est contredit par la possibilité d'adopter des G4 avec seulement 2 tétrades de G empilées (Zhang *et al.*, 2010b). Ensuite, la limite de 7 nt dans les boucles est elle aussi arbitraire, puisque la séquence de l'ARN satellite CEBP5 a été démontrée pour former un G4 avec une boucle composée de 9 nt (Amrane *et al.*, 2012). Des études sur la stabilité de G4 d'ADN de séquences prédéterminées ont démontré que lorsque la première et la troisième boucle sont limitées à 1

seul nucléotide, la boucle centrale pouvait être beaucoup plus longue, soit jusqu'à 30 nt sans que la stabilité du G4 ne soit compromise (Amrane *et al.*, 2012 ; Guédin *et al.*, 2010). Les boucles peuvent donc être beaucoup plus longues que celles décrites par le motif consensus.

La nécessité de posséder des séries de G consécutifs a aussi été remise en question suite à la démonstration que des G4 avec renflement pouvaient se former. C'est-à-dire que même si une série de G est interrompue par la présence d'un autre nucléotide, celui-ci sera extrudé vers l'extérieur de la structure et les tétrades pourront tout de même se former (Mukundan et Phan, 2013). Des G4 avec des renflements peuvent être aussi reconnus par la nucleolin, une protéine liant les G4, que par un anticorps G4 spécifique (Das *et al.*, 2016). De même, il n'y a pas que les guanines qui peuvent former des appariements planaires. Des « mésappariements » de tétrades sont possibles. Cela forme des tétrades hétérogènes dans lesquelles d'autres bases que des guanines, telles que des adénines, des cytosines ou des uraciles peuvent venir compléter la tétrade (Malgowska *et al.*, 2014, 2016 ; Tomasko *et al.*, 2009). Ces tétrades hétérogènes sont évidemment moins stables que les tétrades de G, mais peuvent tout de même être stabilisées lorsqu'elles sont empilées sur elles (Gros *et al.*, 2007). Des études *in vitro* ont aussi démontré que dans certaines conditions, des sites pouvaient être laissés « vacants » dans une tétrade, donc être une « triade » de G  et tout de même permettre la formation d'un G4 en présence de deux autres tétrades de G complètes (Heddi *et al.*, 2016 ; Li *et al.*, 2015). Les séquences permettant les G4 atypiques sont aussi retrouvées dans des régions régulatrices du génome et peuvent aussi posséder des fonctions biologiques (Varizhuk *et al.*, 2017).

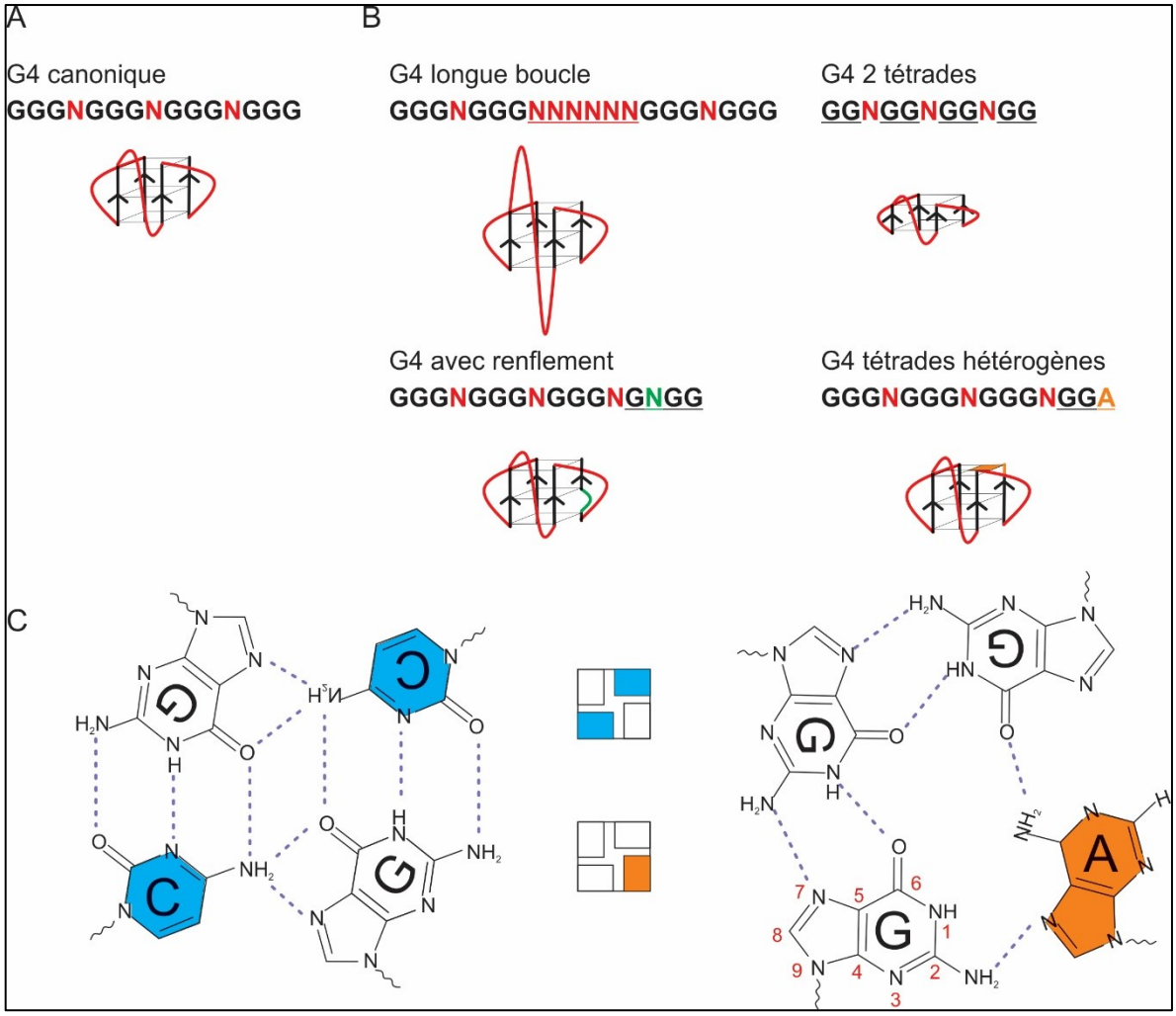



**Figure 9** – Exemples de G4 non canoniques

(A) Description du motif consensus permettant le repliement d'un G4 canonique. (B) Quatre exemples de G4 non canoniques avec les motifs nucléotidiques qui permettent leur formation. (C) Représentations de deux types de tétrades hétérogènes.

Suite à la description de ces nombreux exemples atypiques, on constate que les outils de prédiction de G4 actuels sont à la fois basés sur des définitions trop strictes et donc pas assez sensibles afin d'identifier ces G4 plus variés. À l'inverse, les outils de prédictions sont aussi trop peu spécifiques, car plusieurs des régions à haute densité en séries de G, bien que respectant le motif consensus ne forment pas la structure lorsque testées expérimentalement (Beaudoin et Perreault, 2010).

Dans la majorité des outils de prédiction, les G4 d'ADN et d'ARN sont considérés comme équivalent, et obéissant aux mêmes règles de formation, alors que l'on sait qu'expérimentalement et tel que décrit précédemment que ces deux types de structures

diffèrent. Les particularités des rG4 ne sont pas considérées à part entière dans les méthodes de prédictions par recherche de motifs. Dans la méthode de prédiction de G4 de *RNAfold* qui utilise spécifiquement la prédiction de structures secondaires d'ARN, le modèle thermodynamique utilisé a été construit sur la base de molécules d'ADN et est très limité puisqu'il ne permet aucune séquence atypique.

## Méthodes expérimentales d'évaluation des G4

### Méthodes *in vitro*

Afin de valider si toute prédiction de repliement G4 est valable, il est essentiel de confirmer expérimentalement si la structure est formée. Grâce aux particularités uniques des G4 et rG4 énumérées précédemment, de nombreuses techniques *in vitro* ont été développées afin de déterminer si un repliement G4 est présent et quels nucléotides exacts d'une séquence sont impliqués dans la structure. Aucune de ces techniques n'est complète en soi, et chacune est informative sur certains aspects, mais limitée sur d'autres. Afin d'avoir une preuve formelle de repliement G4 et une caractérisation complète, il est nécessaire de combiner plusieurs de ces techniques.

*Cristallographie et RMN*

La méthode par excellence afin de déterminer la structure de toute biomolécule est la cristallographie. Cela consiste à réunir les conditions nécessaires : pression, pH, concentration en sels, présence d'agents cristallisants, concentration de la molécule d'intérêt, etc. afin d'obtenir un cristal. Celui-ci est ensuite soumis à des rayons X et le patron de diffraction de ces rayons sera analysé afin de déterminer la position de chacun des atomes dans le cristal et ainsi de connaître la structure globale de la biomolécule. Cette méthode permet d'obtenir des structures à très haute résolution. Cependant, c'est une méthode extrêmement fastidieuse qui est dépendante de la capacité à obtenir un cristal unique. De plus, les conditions de cristallisation sont souvent très différentes des conditions biologiques où la molécule biologique effectue sa fonction. Des structures G4 ont été déterminées par cristallographie pour l'ADN et l'ARN des répétitions télomériques (Neidle et Parkinson, 2008). La cristallographie a aussi permis de déterminer le site de reconnaissance de la protéine FMRP qui reconnaît la jonction duplex-quadruplex sur des ARNm (Vasilyev *et al.*, 2015).

La résonance magnétique nucléaire (RMN) est aussi une technique de résolution structurale qui peut s'appliquer à plusieurs types de biomolécules. Cette méthode est basée sur l'excitation et l'entrée en résonance du spin des protons du noyau des atomes selon leur position dans un champ magnétique et leurs interactions avec la densité électronique des autres atomes à proximité. Cela varie selon la structure adoptée par la molécule. L'analyse des déplacements chimiques résultant de la résonance des protons permet de déterminer la structure globale de la molécule. Tout comme la cristallographie, cette technique permet d'élucider des structures complètes avec une excellente résolution. Son avantage est qu'elle s'effectue avec des molécules en solution et donc peut permettre de voir diverses conformations dynamiques de la structure plutôt qu'un cristal fixe. Cela est plus près des conditions biologiques dans lesquelles les structures se retrouvent, mais dans le cas des acides nucléiques, requiert tout de même de grandes concentrations. De plus, pour limiter la complexité d'attribution des positions des atomes, on n'utilise que de courtes séquences d'acides nucléiques. Dans le cas des G4 en particulier, plusieurs structures ont été déterminées grâce à cette technique (Adrian *et al.*, 2012). Récemment, des avancées permettent d'évaluer la structure rG4 par RMN *in cellulo* ont aussi été réalisées (Bao et Xu, 2018).

Sans nécessairement résoudre toute la structure ni attribuer tous les pics de déplacement de chaque atome, la présence de G4 et de rG4 peut être confirmée grâce à la présence de pics de déplacements chimiques caractéristiques des protons iminos des guanines dans les tétrades. Ces déplacements se situent entre 10 et 12 ppm. Les structures secondaires Watson-Crick quant à elles donnent des déplacements chimiques situés dans la zone de 13 à 14 ppm. En augmentant graduellement la concentration de potassium, l'apparition graduelle des pics des protons iminos (un pic pour chaque guanine impliquée dans une tétrade) constitue une preuve de formation *in vitro* de G4.

*Dichroïsme circulaire et dénaturation thermique*

Les caractéristiques uniques des G4 permettent aussi de les distinguer et de les caractériser grâce à des techniques spectroscopiques. La première qui est la plus courante est le dichroïsme circulaire (DC). Cette méthode mesure l'absorption différentielle de la lumière circulaire polarisée selon la structure des molécules. Celle-ci est rapportée en mesure d'ellipticité ($\theta$). Les structures G4 peuvent être reconnues par des spectres très spécifiques

en DC. Les G4 et rG4 de topologie parallèle sont reconnus par la présence d'un creux à 245 nm et d'un pic à 264 nm qui forment ce spectre caractéristique. Les G4 antiparallèles présentent un creux à 260 nm et un pic à 295 nm et les hybrides ont deux pics à 295 nm et 260 nm et un creux à 245 nm (Del Villar-Guerra *et al.*, 2018). Il est possible de monitorer la présence du pic spécifique de DC à 264 nm des G4 parallèles pour mesurer la dénaturation thermique de la structure et ainsi déterminer sa température de dénaturation ($T_m$) et évaluer sa stabilité. Pour ce faire, on mesure le pic de DC à 264 ou 295 nm selon la topologie de la molécule G4 repliée. En augmentant graduellement la température, on voit ainsi le pic disparaître lorsque la structure est dénaturée par la chaleur. Les G4 et rG4 étant très stables, ils sont caractérisés par des $T_m$ très élevées en présence de $K^+$ essentiel à leur formation comparativement à la présence de $Li^+$ qui est souvent utilisé comme contrôle négatif.

Ces techniques spectroscopiques permettent de déterminer si la structure secondaire de la séquence d'acides nucléiques forme ou non un G4. Elle permet de déterminer sa topologie (parallèle, antiparallèle ou mixte) selon le spectre obtenu et permet de mesurer une de ces caractéristiques propres, soit le $T_m$ de la structure. Cependant, cette technique nécessite d'utiliser de grandes concentrations d'acides nucléiques afin d'obtenir un signal suffisant (en µM), donc en quantité beaucoup plus grande que ce qui est probable en condition physiologique. De plus, cela favorise la formation de structure intermoléculaire. Les structures secondaires Watson-Crick d'ARN donnent des signaux de DC similaires, avec aussi des pics autour de 260 nm (Fay *et al.*, 2017). Afin de ne pas embrouiller le spectre de DC du G4 par la présence de plusieurs structures secondaires, uniquement de courtes séquences peuvent être utilisées (uniquement le motif G4 sans les séquences adjacentes). Finalement, bien que le DC et la dénaturation thermique permettent de déterminer si la séquence forme ou non un G4, cela ne permet pas de savoir quels nucléotides exactement sont impliqués dans les tétrades ou dans les boucles.

*Fluorescence*

La structure G4 peut être étudiée grâce à la spectroscopie utilisant la lumière visible, mais aussi grâce à la spectroscopie à fluorescence. Plusieurs outils ont été développés afin de déterminer la structure secondaire des acides nucléiques et peuvent être adaptés pour confirmer la présence de structure G4. Ces techniques utilisent la présence de bases modifiées ou l'ajout de fluorophores attachés à la séquence d'acides nucléiques. Par exemple, l'adénine

modifiée en 2-aminopurine émet de la fluorescence lorsqu'elle est dans une position libre (non appariée) et n'en émet plus lorsqu'elle est impliquée dans une paire de bases ou camouflée dans une structure tertiaire. C'est une propriété qui peut être utilisée pour mesurer le repliement G4 en incorporant l'aminopurine dans une boucle (Gray *et al.*, 2010). Une autre technique qui peut être utilisée est le FRET (*Förster Resonance energy transfer).* Avec cette technique, une émission de fluorescence pourra être mesurée uniquement si un fluorophore donneur se trouve à très grande proximité d'un fluorophore accepteur. En excitant le fluorophore donneur, celui-ci émettra de la fluorescence à une longueur d'onde spécifique. Si les deux fluorophores se retrouvent à proximité, l'énergie émise à une certaine longueur d'onde par le premier fluorophore permettra d'exciter le fluorophore accepteur. Celui-ci émettra alors à son tour à une longueur d'onde différente. C'est cette seconde émission qui est mesurée et qui survient seulement si les deux fluorophores se retrouvent assez près l'un de l'autre dans la structure pour permettre le FRET. En connaissant la séquence et en prédisant les structures secondaires possibles d'une séquence d'intérêt, il est possible de créer des designs différents de séquences dans lesquels on peut inclure des acides nucléiques modifiés ou des fluorophores donneurs et accepteurs permettant de mesurer de l'émission de fluorescence ou de FRET uniquement si un G4 est replié par exemple, donc en comparant en conditions avec ou sans $K^+$ (Swiatkowska *et al.*, 2016 ; Ying *et al.*, 2003).

Fait intéressant, les G4 d'ADN et d'ARN ont la capacité intrinsèque d'émettre de la fluorescence, et ce sans modification de leur séquence. Cette propriété des G4 est due à la conjugaison étendue dans les tétrades. Par contre, la capacité d'émission varie grandement selon le nombre de tétrades empilées, la taille ainsi que la composition des boucles (Kwok *et al.*, 2013). Malgré cette limite, on peut tout de même utiliser la présence de la tétrade de G non modifiée pour étudier le repliement par fluorescence. Cet « anneau » planaire peut être reconnu et lié spécifiquement par les petites molécules aromatiques. C'est le cas entre autres de la Thioflavine T (ThT) de la N-Méthyl Mésoporphyrine (NMM) (Renaud de la Faverie *et al.*, 2014 ; Sabharwal *et al.*, 2014). Ces molécules seules en solution n'émettent pas de fluorescence. Par contre, une fois empilée sur une tétrade, la formation de l'empilement π et la conjugaison entraînent une forte émission de fluorescence lorsque le ligand est excité à une longueur d'onde particulière. Cette propriété peut donc être utilisée en solution en présence d'une séquence G4 et du ligand fluorescent, encore une fois en comparant des

conditions favorables ou non à la formation de G4. De la fluorescence élevée sera mesurée uniquement si le G4 est replié. Il existe toute une gamme de ligands fluorescents spécifiques aux G4 (Bhasikuttan et Mohanty, 2015 ; Vummidi *et al.*, 2013). Certains reconnaissent autant les G4 d'ADN que d'ARN alors que d'autres sont spécifiques à l'une ou l'autre des molécules. Le ligand CyT par exemple n'émet de la fluorescence qu'en présence de rG4 et non en présence d'équivalents d'ADN (Xu *et al.*, 2015). Certains ligands fluorescents peuvent aussi discriminer entre des G4 de topologies différentes soit parallèle ou antiparallèle. Le NMM reconnaît préférentiellement les G4 de topologie parallèle tandis que le cyanovinyl-pyridinium triphénylamine (CPT) reconnaît les G4 antiparallèles (Lai *et al.*, 2014 ; Sabharwal *et al.*, 2014).

Les limites de ces techniques sont qu'elles nécessitent encore une fois de grandes concentrations d'acides nucléiques afin d'obtenir un signal clair, ce qui peut entraîner la formation de structures intermoléculaires. Pour l'utilisation du FRET et des nucléotides modifiés fluorescents, le succès de la technique dépend de la connaissance préalable des structures secondaires possibles. De plus, elle est limitée par les designs expérimentaux possibles selon la composition et la longueur de la séquence d'intérêt. L'utilisation de ligands fluorescents est une méthode simple et rapide afin d'évaluer plusieurs séquences PG4. Par contre, la sensibilité de cette technique peut être affectée si une même séquence adopte plusieurs conformations G4 différentes en solution ou encore en présence d'un G4 non canonique, deux situations où l'affinité du ligand envers le G4 pourrait être affectée et donc diminuer l'émission de fluorescence et entraîner un faux négatif. À l'inverse, la présence du ligand pourrait venir stabiliser des tétrades en s'empilant sur elles et ainsi entraîner la stabilisation d'une séquence en structure G4 qui ne serait pas formée si la séquence était seule en solution, ce qui constitue un faux positif.

*Retardement sur gel*

Dans un gel de polyacrylamide (PAGE) natif, les séquences d'acides nucléiques ont des mobilités électrophorétiques (vitesse de migration) différentes dues à leur poids moléculaire, leur taille et leur forme ou leur structure adoptée (Sun et Hurley, 2010). Les séquences repliées migreront de façons différentes que des séquences non appariées dans un essai EMSA (*Electrophoretic mobility shift assay)*. De même, les G4 intermoléculaires migreront plus lentement que les G4 intramoléculaires à cause du plus haut poids moléculaire des

premiers. Les rG4 avec leur topologie parallèle peuvent être très compacts et migrer sans retardement sur un gel natif comparativement à des G4 de topologies antiparallèles. On peut donc comparer les retards et les changements dans la migration de séquences G4 avec une séquence contrôle dans lesquelles on abolit le potentiel G4 par des mutations dans les séries de G ou encore une fois en comparant en présence d'ions $K^+$ ou $Li^+$. Il est même possible de tremper ces gels après migration dans des solutions contenant des ligands fluorescents tel que le NMM mentionné précédemment pour observer quelles bandes de mobilités différentes sont fluorescentes et donc identifier celles qui correspondent à la formation de G4 ou de rG4.

*Cartographie*

Pour l'élucidation des structures secondaires d'ARN, de multiples techniques de cartographie ont été développées. L'ensemble de ces méthodes reposent sur un principe commun. Il s'agit de marquer avec un phosphate radioactif ou un fluorophore une des extrémités, soit 5' ou 3' d'un brin d'ARN (souvent synthétisé par transcription *in vitro*) ou d'une amorce d'oligonucléotides. Par la suite, deux alternatives sont possibles. Avec la première, on entraîne le clivage partiel des brins en solution à des positions spécifiques selon l'utilisation de ribonucléases (RNase) qui clive spécifiquement certaines structures secondaires de façon préférentielle : par exemple, la RNase V1 clive les ARN double-brins, alors que la RNase T1 clive le lien adjacent au 3'-phosphate des G non appariés et la RNase T2 celui des A non appariés (Mailler *et al.*, 2018). L'autre alternative consiste à utiliser des agents chimiques qui vont réagir de façon covalente avec certains nucléotides libres (non appariés) et ainsi ajouter des adduits encombrants sur la molécule qui viendront bloquer une transcriptase inverse et donc arrêter l'élongation du segment d'ADNc lors de la transcription inverse à partir de l'amorce marquée du départ. Suite à l'une ou l'autre de ces alternatives, on se retrouve en solution avec différents brins d'ARN ou d'ADNc de longueurs différentes selon où le clivage ou l'arrêt de transcription inverse a eu lieu. On sépare ces fragments à l'aide de gels de polyacrylamide dénaturants, aussi appelés gels de séquençage, avec lesquels on fait co-migrer une séquence qui sert d'échelle moléculaire afin de pouvoir identifier les positions de chaque nucléotide. En comparant les patrons de migration résultant du traitement avec les diverses RNases et des diverses conditions en solution, on peut déduire la structure secondaire adoptée par la séquence initiale. Différentes techniques de cartographie ont donc été adaptées afin de déterminer le repliement rG4 et les nucléotides impliqués.

Arrêt de la transcriptase inverse (*RTS, reverse transcriptase stalling assay*)

Cette méthode de cartographie est l'une des premières méthodes utilisées pour l'identification des séquences formant un rG4 et l'une des plus simples. Elle est basée sur la grande stabilité de la structure. En effet, les rG4 sont tellement stables qu'ils peuvent arrêter l'élongation d'une enzyme transcriptase inverse. Suite à l'hybridation d'une amorce en amont de la séquence PG4, la transcription inverse sera bloquée lors de la rencontre du dernier « G » en 3' du motif PG4. Cependant, il n'y a pas que les structures secondaires rG4 qui peuvent entraîner le décrochage ou des pauses de la transcriptase inverse, il est donc important encore une fois de comparer les patrons d'arrêt en condition favorable ($K^+$) et défavorable ($Li^+$, séries de G4 mutées) aux rG4. Lorsque le rG4 n'est pas formé, la transcription inverse devrait être complète et montrer peu d'arrêts ou des arrêts de faible intensité, alors qu'en condition favorable au rG4 les arrêts seront situés principalement aux derniers G de chaque séries de G impliquées dans les tétrades du rG4 (Kumari *et al.*, 2015 ; Kwok et Balasubramanian, 2015).

Protection RNase T1

La RNase T1 a une spécificité de clivage pour les G non appariés. Cela s'avère particulièrement utile dans l'étude des rG4. En condition favorable, en présence de $K^+$, les G impliqués dans les tétrades vont s'apparier entre eux et seront donc protégés du clivage par la RNase T1. Tandis qu'en condition défavorable, en présence de $Li^+$, les G non appariés seront clivés. En comparant ces deux patrons de clivage, on peut donc déterminer que les G protégés sont ceux impliqués dans un quadruplex. Néanmoins, les G formant des paires de bases G-C sont aussi protégés du clivage par la RNase T1. Donc, si en condition $Li^+$ certains G forment plutôt des appariements G-C, ceux-ci demeureront protégés. Il est aussi important de noter que les RNase ont des « préférences » de site de clivage, mais qu'elles sont très processives et ne sont donc pas hautement spécifiques. Elles peuvent ultimement cliver un brin d'ARN un peu partout. La cartographie de protection à la RNase T1 ne donne des informations que pour les nucléotides G et doit souvent être utilisée en combinaison avec d'autres RNases si on veut obtenir une structure secondaire plus détaillée, par exemple pour connaître les structures secondaires adjacentes ou celles des boucles.

Cartographie au DMS

Le dimethylsulfate (DMS) est un composé chimique qui peut réagir avec les guanines pour méthyler la position N7 de la base azotée. Il peut aussi méthyler la position N1 des adénines et N3 des cytosines. Cette méthylation peut survenir uniquement si ces positions sont accessibles dans le solvant, donc les nucléotides seront protégés s'ils sont impliqués dans des ponts hydrogène et des structures secondaires. La présence du méthyl sur les bases A et C inhibera la transcriptase inverse lors d'une réaction d'extension d'amorce et permettra de détecter lesquels sont accessibles ou non. Afin de détecter les G méthylés, il faudra une étape supplémentaire, soit traiter la séquence avec de l'aniline. Cela entraînera le clivage du brin d'ARN aux positions des G méthylés (Wells *et al.*, 2000). Dans le cas des rG4, la position N7 est inaccessible à la méthylation puisqu'elle est protégée par la formation du lien Hoogsteen dans la tétrade (Sun et Hurley, 2010). On peut donc déduire quels sont les G impliqués dans le rG4 en voyant lesquels sont protégés du clivage. Puisque le DMS est un composé de très petite taille, il peut traverser la barrière phospholipidique cellulaire. Ainsi, il est aussi possible d'utiliser le DMS pour effectuer la cartographie de structures secondaires *in cellulo* (Guo et Bartel, 2016).

Cartographie *in-line*

La méthode de cartographie *in-line* diffère légèrement des autres méthodes puisqu'elle ne nécessite pas l'utilisation de RNase ou d'agent chimique. Cette méthode est possible grâce à la capacité intrinsèque de l'ARN à s'autocliver. C'est entre autres pour cette raison que l'ARN se dégrade « facilement » et rapidement. En effet, la présence du groupement hydroxyle en position 2' du ribose (2'-OH), en condition basique et en présence de magnésium ($Mg^{2+}$), permet l'attaque du lien phosphodiester et le réarrangement électronique résultant en un lien cyclique 2',3'-phosphate sur le ribose responsable de l'attaque et donc du clivage du squelette (**Figure 10**).

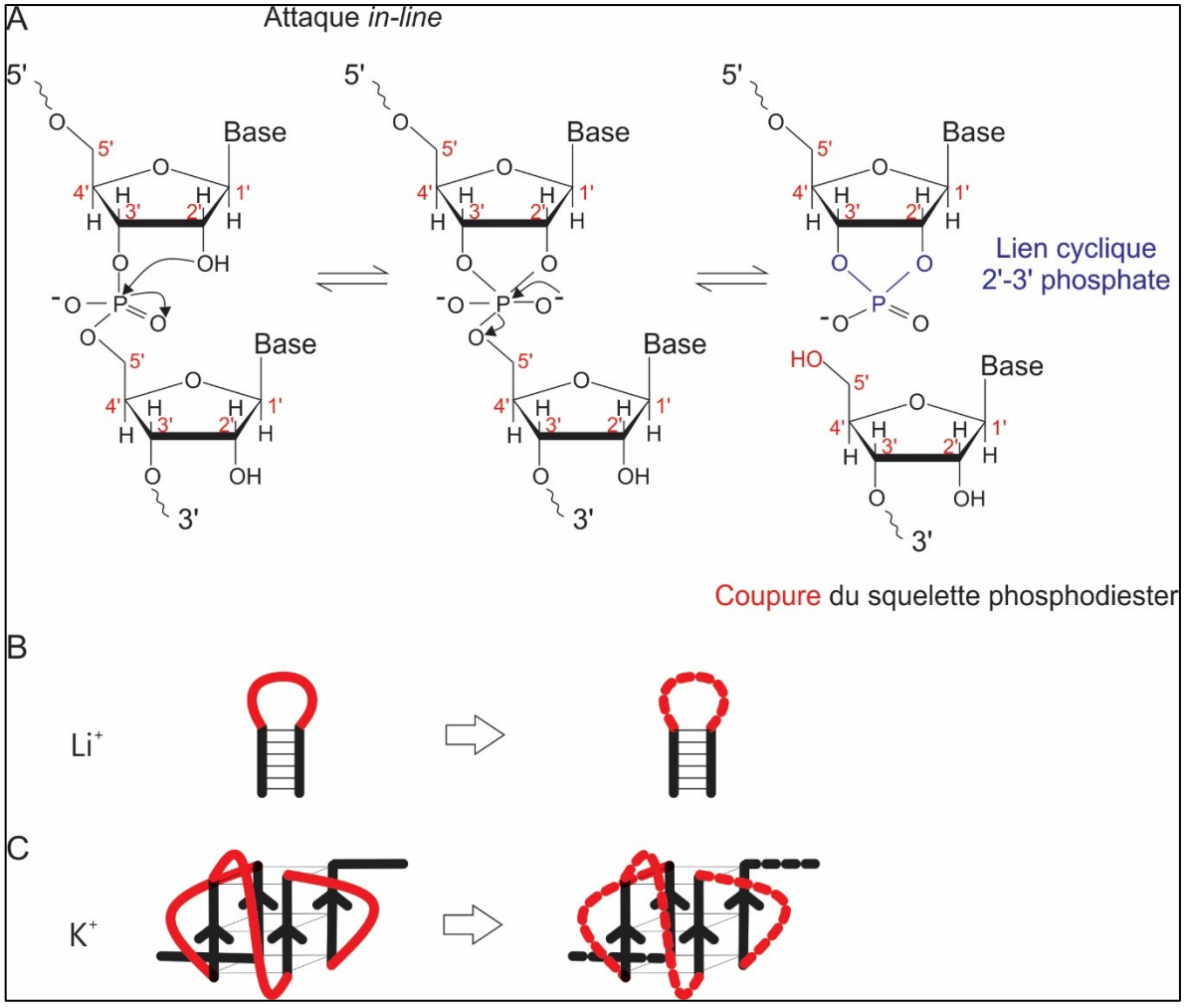



**Figure 10** – Attaque *in-line*

(A) Représentation de la réaction de coupure *in-line* pouvant se produire sur un brin d'ARN entraînant son auto-coupure. (B) Les sites préférentiels de coupure sur une structure secondaire canonique d'ARN sont les boucles. (C) Tandis que pour un rG4, les sites préférentiels sont les boucles reliant les tétrades et les nucléotides adjacents au rG4 en 5' et en 3'.

Cette réaction spontanée survient préférentiellement lorsque le squelette phosphodiester se retrouve dans cette conformation particulière « en ligne » d'où le nom d'attaque *in-line*. Les régions du squelette qui sont plus flexibles sont donc plus susceptibles d'adopter cette conformation et d'être clivées. Les régions les plus flexibles sont les régions où les nucléotides sont non appariés.

Il est donc possible d'utiliser cette propriété de l'ARN à notre avantage pour la détermination de sa structure secondaire, les régions structurées étant moins susceptibles au clivage. L'idée consiste à comparer les patrons de clivage spontané de l'ARN dans des

conditions différentes. Si un rG4 est formé, on s'attend à ce que les régions les plus flexibles de la structure secondaire soient les régions non appariées : les 3 boucles ainsi que les nucléotides immédiatement avant et après le rG4. Tandis que dans une structure secondaire d'ARN canonique, ce seront les boucles des tiges-boucles et les boucles internes qui seront clivées préférentiellement. Encore une fois, pour l'étude des rG4 on peut comparer le patron de clivage entre la condition $K^+$ favorable et $Li^+$ défavorable. Si en condition $K^+$ on observe un clivage supérieur pour les nucléotides situés entre les séries de G prédites du rG4, alors que ce clivage n'est pas observé en condition $Li^+$, on peut conclure à la formation d'un rG4.

**Méthodes in cellulo**

Les méthodes de détection et d'élucidation des structures G4 *in vitro* sont utiles afin de déterminer de façon précise les particularités d'une structure d'intérêt déjà choisie et de déterminer ses caractéristiques précises. Par contre, ces expériences sont souvent loin de répliquer les conditions biologiques réelles dans lesquelles les G4 sont modulés et dans lesquelles ils peuvent jouer leur rôle, d'où l'importance d'utiliser des techniques de détection de la structure *in cellulo.* La formation et la présence de G4 d'ADN et d'ARN en cellule ont été confirmées de multiples façons.

*Anticorps spécifiques aux structures G4*

La toute première preuve de la formation de structure G4 d'ADN en cellule a été obtenue grâce à l'immunofluorescence. Un anticorps reconnaissant spécifiquement les structures G4 comparativement aux autres structures secondaires d'ADN a été utilisé. Il a permis la détection des G4 d'ADN dans l'organisme cilié *Stylonychia Lemnae,* particulier pour son macronoyau avec un nombre élevé de télomères de séquences répétées TTTTGGGG (Schaffitzel *et al.*, 2001). Cette étude a démontré que les séquences télomériques pouvaient bel et bien adopter des structures G4 *in cellulo*.

Par la suite, un autre anticorps spécifique au G4 appelé BG4 a été développé par la technologie d'expression phagique (*phage display*) afin de reconnaître les G4 avec une haute affinité dans des lignées cellulaires U2OS (ostéosarcome), HeLa (carcinome cervical), HT1080 (fibrosarcome), MCF-7 (adénocarcinome mammaire) et MDA-MB-231 (carcinome mammaire) (Biffi *et al.*, 2013). On a pu observer que les signaux ce situaient principalement au noyau, aux extrémités des chromosomes et qu'il y en avait plus lors de la phase S de

réplication (où les deux brins complémentaires d'ADN sont séparés). Des ligands stabilisateurs de G4 ont aussi été utilisés et le signal de l'anticorps spécifique augmentait dans cette condition. Des résultats équivalents ont été obtenus dans des cellules humaines et murines pour un second anticorps monoclonal spécifique au G4 d'ADN appelé 1H6 (Henderson *et al.*, 2014). Par contre, la prudence est requise dans l'interprétation de ces résultats. Récemment, on a démontré que cet anticorps semble plutôt reconnaître spécifiquement des séquences de poly(T) dont la structure est restreinte par une structure adjacente comme un G4 et non le G4 lui-même (Kazemier *et al.*, 2017).

De façon très intéressante, on a aussi constaté que l'anticorps BG4 permettait de visualiser un signal dans le cytoplasme des cellules, et que ce signal était toujours présent suite à un traitement à la désoxyribonucléase (DNase). Vraisemblablement, cet anticorps développé pour reconnaître des G4 d'ADN de topologie parallèle permet aussi de détecter en cellule la présence de rG4. Ce qui constitue un argument très important pour l'importance biologique des structures rG4 dans les ARNm (Biffi *et al.*, 2014a). Il a aussi été démontré par immunohistochimie avec cet anticorps que les G4 étaient plus abondants dans les cellules cancéreuses du foie et de l'estomac que dans des tissus non transformés (Biffi *et al.*, 2014b).

*Sondes*

Outre l'utilisation d'anticorps spécifiques, plusieurs groupes de recherche ont développé des molécules chimiques ayant des affinités spécifiques pour les G4 et rG4. Ces sondes peuvent être utilisées en cellule en combinaison avec un fluorophore afin de détecter la présence de G4 par microscopie. Avec ces méthodes, les G4 et rG4 sont aussi observés dans le noyau et le cytoplasme de cellules vivantes confirmant de nouveau leur présence intracellulaire et leur repliement dynamique (Amor *et al.*, 2017 ; Chen *et al.*, 2018b ; Laguerre *et al.*, 2015, 2016 ; Manna et Srivatsan, 2018).

*Gènes rapporteurs*

La façon la plus courante d'évaluer si une séquence PG4 peut se former *in cellulo* et avoir un effet sur l'expression d'un transcrit est l'utilisation d'un gène rapporteur. Dans ce type d'essai, la séquence PG4 d'intérêt est insérée dans un plasmide codant pour un gène rapporteur facilement détectable, que ce soit un gène de luciférase (Fluc ou Rluc), chloramphénicol acétyltransferase (CAT) ou β-galactosidase (β-gal) (Halder *et al.*, 2012). En

parallèle, on insère dans le même vecteur rapporteur une séquence semblable à la séquence PG4 d'intérêt, mais dans laquelle le PG4 est soit complètement enlevé ou muté pour empêcher sa formation. Afin d'abolir la formation d'un rG4, une première mutation possible est d'éliminer complètement une série de G de la séquence initiale afin qu'il n'en reste moins que 4, le nombre minimal requis, dans la séquence mutée. Une seconde mutation possible est de substituer quelques G par un autre nucléotide dans les séries, par exemple en mutant la série GGG en GAG, afin d'éliminer la possibilité de former des tétrades. Puisque plusieurs G4 atypiques peuvent se former malgré des renflements ou en utilisant des séries de G supplémentaires qui pourraient être présentes une fois la séquence insérée dans le plasmide, il est important de vérifier préalablement *in vitro* que le mutant de rG4 dessiné abolit bel et bien la formation du rG4. Selon la position naturelle du rG4 dans son transcrit d'origine on peut l'insérer dans le 5'UTR ou le 3'UTR du gène rapporteur.

Par la suite, ces gènes rapporteurs avec la séquence PG4 sauvage (wt) et son contrôle rG4-muté (mut) sont transfectés individuellement en parallèle dans une lignée cellulaire choisie. L'intensité de l'expression mesurée, comme la bioluminescence dans le cas d'un gène rapporteur luciférase, sera proportionnelle à l'expression du gène. En comparant les constructions wt et mutées, il sera possible de savoir si la présence du rG4 affecte l'expression.

**Ciblage des G-quadruplexes**

Plusieurs des méthodes expérimentales utilisées pour la détection et la caractérisation du repliement des G4 dépendent de la reconnaissance spécifique de la structure. De plus, afin de mieux comprendre leurs fonctions biologiques il peut être intéressant d'être capable de contrôler leur repliement en le forçant ou en l'empêchant. Pour ce faire, des outils permettant de cibler spécifiquement la structure G4 sont nécessaires, certains ont été brièvement présentés dans les sections précédentes.

*Ligands*

Les ligands fluorescents permettant de détecter les G4 ont été abordés précédemment. Néanmoins, dès que des motifs G4 ont été prédits dans les séquences télomériques et observés comme étant d'excellents inhibiteurs de la télomérase, une enzyme surexprimée dans les cellules cancéreuses, une course s'est engagée afin d'identifier des ligands

pharmacologiques pouvant cibler et stabiliser les G4. À ce jour, la majorité des ligands connus ont été sélectionnés et optimisés afin de reconnaître la séquence des répétitions télomériques, ainsi que les G4 présents dans quelques promoteurs de proto-oncogènes (par exemple c-MYC et N-RAS), tous étant constitués d'une séquence ADN respectant le motif canonique des G4. Très tôt, on a constaté l'importance de bien caractériser les différences structurales entre les différents G4 puisque cela affectait la liaison avec les ligands. Par exemple, le promoteur du gène c-KIT possède 2 structures G4 repliées qui ne sont pas stabilisées de façons identiques selon les ligands (Neidle, 2012). À cause de toutes les menues caractéristiques possibles des G4 énumérées dans les sections précédentes, il n'existe pas de ligand G-quadruplex « universel ».

Une banque de données existe regroupant les ligands de G4 répertoriés dans la littérature. Celle-ci regroupe plus de 800 composés développés ciblant les G4 d'ADN principalement et quelques-uns spécifiques aux rG4 (Li *et al.*, 2013). Les ligands sont classés en différentes familles selon les groupements chimiques formant leur cœur : les acridines (ex. : BRACO-19)(Read *et al.*, 2001), les pyridostatines (ex. : PDS)(Rodriguez *et al.*, 2008), les bisquinoliniums (ex. : Phen-DC3)(De Cian *et al.*, 2007), les porphyrines (ex. : NMM, TmPyP4) (Izbicka *et al.*, 1999), etc. Malgré leurs différences de groupements fonctionnels, leurs modes de reconnaissance et d'interaction avec les G4 sont presque tous identiques. Ils sont basés sur l'affinité des groupements polyaromatiques de ces molécules permettant leur empilement π et les interactions électrostatiques sur la surface de la tétrade, ce qui favorise ainsi le repliement et la stabilité de la structure. Il existe une autre classe de ligands, beaucoup plus limitée, qui reconnaissent quant à eux les sillons des G4 ADN (Di Leva *et al.*, 2014). La molécule la plus connue de cette catégorie est la distamycin A (Martino *et al.*, 2007). La grande majorité des ligands stabilise les G4, mais il en existe aussi qui les déstabilise comme le triarylpyridine (TAP)(Waller *et al.*, 2009).

Certains ligands ciblent des G4 de topologie parallèle et peuvent donc reconnaître autant les G4 d'ADN que d'ARN qui adoptent cette topologie. C'est le cas pour les ligands PDS et Phen-DC3. Le ligand TmPyP4 est particulier à cause de ses effets opposés sur les G4 d'ADN et d'ARN. Il est stabilisateur pour les premiers, alors qu'il est déstabilisateur pour les seconds (Morris *et al.*, 2012 ; Zamiri *et al.*, 2014). Des ligands ont été modifiés et développés pour être spécifiques envers les G4 d'ARN uniquement. Particulièrement, une

variation du PDS appelée carboxy-PDS (cPDS) (Di Antonio *et al.*, 2012 ; Rocca *et al.*, 2017) et un composé appelé N-TASQ (Laguerre *et al.*, 2015). Ces deux composés ont l'avantage de permettre le ciblage des rG4 *in cellulo* (Kwok et Balasubramanian, 2015 ; Yang *et al.*, 2018).

*Oligonucléotides antisens (ASO)*

Les ligands sont d'excellents outils, mais ils peuvent reconnaître et cibler simultanément plusieurs G4 situés sur plusieurs gènes ou ARNm présents dans la cellule. Tel qu'abordé précédemment, une façon beaucoup plus spécifique de cibler les G4 est l'utilisation de nucléotide antisens (ASO) qui peuvent venir se lier par appariement de séquence et ainsi déstabiliser le G4 en compétitionnant pour former des paires de bases canoniques, ou à l'inverse, favoriser le G4 en séquestrant les nucléotides compétitifs adjacents. Cette méthode a été démontrée pour favoriser ou empêcher la formation de rG4 dans le 5'UTR d'un ARNm (Rouleau *et al.*, 2015). Par contre, cette méthode a aussi des limites. En plus des défis posés pour s'assurer de la spécificité de liaison des ASO, de leur entrée en cellule, de l'évitement de leur dégradation et de leur distribution en cellule, soit de rejoindre le noyau, certains G4 et rG4 une fois formés sont tellement stables que l'ASO ne peut compétitionner pour défaire la structure.

## Rôles biologiques des G4

Les quadruplexes sont des structures secondaires non canoniques très stables, mais dynamiques, dont la formation et la stabilité sont modulées à la fois par leur propre séquence et par des facteurs environnementaux. Ils sont prédits dans des régions régulatrices du génome et du transcriptome et il est possible de les détecter et de les cibler dans des cellules. Quels sont donc leurs rôles et leurs impacts biologiques ?

### Présence des G4 chez plusieurs organismes

Tel que mentionné dans la section précédente concernant la prédiction des PG4, il y a une forte prévalence de quadruplex dans le génome et le transcriptome humain, particulièrement dans des régions régulatrices. Ceux-ci sont également retrouvés, et certains conservés, dans des régions régulatrices chez plusieurs autres espèces de mammifères et autres branches des eucaryotes (Verma *et al.*, 2008). Par conservation, on entend que la séquence d'acide

nucléique est identique ou similaire et qu'elle est présente dans la même localisation du génome chez des espèces différentes d'un point de vue évolutif. Cela indique que la séquence formant le motif PG4 a été maintenue lors de la sélection naturelle. L'homologie des séquences et leur localisation sont observées par alignements multiples des génomes et par la prédiction de motifs PG4 (Kikin *et al.*, 2006 ; Yadav *et al.*, 2008 ; Frees *et al.*, 2014 ; Dhapola et Chowdhury, 2016 ; Marsico *et al.*, 2019). Des G4 sont prédits et caractérisés chez les levures, ainsi que la présence d'hélicases qui les reconnaissent (Capra *et al.*, 2010 ; Sabouri *et al.*, 2014). Les G4 sont retrouvés de façon non aléatoire dans les génomes de plusieurs micro-organismes (bactéries, archéobactéries, parasites)(Ding *et al.*, 2018 ; Leeder *et al.*, 2016 ; Saranathan et Vivekanandan, 2018). Le contrôle de la formation de G4 est même une stratégie pour lutter contre la virulence de certains pathogènes (Harris et Merrick, 2015). Ils sont présents également dans diverses plantes (Garg *et al.*, 2016 ; Mullen *et al.*, 2010). La présence de rG4 dans les ARNm pouvant affecter la régulation de la traduction chez *Arabidopsis thaliana* (Kwok *et al.*, 2015) et affecter le développement vasculaire du phloème (Cho *et al.*, 2018). Les G4 sont conservés dans plusieurs familles virales à génomes ADN et ARN à l'exception des virus à génome double-brin, ce qui semble logique puisque le brin complémentaire compétitionne avec la formation de G4. En utilisant le contenu PG4 uniquement, il est même possible de reclasser les séquences des génomes de virus dans les bonnes familles virales (Lavezzo *et al.*, 2018). En somme, la présence de séquences PG4 conservées dans l'ensemble de l'arbre de la vie constitue un argument de taille concernant leur importance biologique.

**Rôles des G4 d'ADN**

Afin de mieux comprendre le domaine d'étude des G4 d'ARN, il est nécessaire de faire un survol des fonctions biologiques identifiées préalablement pour les G4 d'ADN. Une des fonctions les mieux caractérisées des G4 d'ADN est leur implication dans la maintenance et la synthèse des extrémités télomériques (Lipps et Rhodes, 2009). Les structures G4 sont aussi retrouvées dans les promoteurs de plusieurs oncogènes tels BCL2, VEGF, KRAS, c-MYC, c-KIT et WNT1 où ils ont un effet inhibiteur sur la transcription et la liaison de facteurs de transcription (Kuo *et al.*, 2015a). Les G4 sont aussi associés aux origines de réplications, aux régions d'instabilité génomique, à la régulation épigénétique et à la

recombinaison entre autres des immunoglobulines (Dempsey *et al.*, 1999 ; Hänsel-Hertsch *et al.*, 2017 ; Mao *et al.*, 2018 ; Prioleau, 2017).

### Rôles biologiques des G4 d'ARN

*rG4 et ARN non codants*

Les rG4 peuvent être adoptés autant par les ARN codants que non codants. Ils ont des fonctions différentes selon la classe d'ARN. Cependant, leur effet est majoritairement répressif, la formation de rG4 dans les précurseurs des petits ARN non codants peut empêcher la formation ou la reconnaissance de la structure double-brin reconnue par diverses nucléases que ce soit Drosha, Dicer et des hélicases qui sont responsables de la maturation des petits ARN fonctionnels comme les microARN (miARN) et les piwi-ARN (piARN) (Ghosh *et al.*, 2018 ; Pandey *et al.*, 2015 ; Rouleau *et al.*, 2018 ; Vourekas *et al.*, 2015). La présence de rG4 dans la forme mature de ces petits ARN ou sur l'ARNm à proximité de leur site de liaison affecte aussi leurs fonctions en empêchant ou en facilitant la liaison avec leur ARN cible (Lung Chan *et al.*, 2018 ; Rouleau *et al.*, 2017a ; Stefanovic *et al.*, 2015). Des rG4 sont aussi formés dans d'autres familles d'ARN non codants tels que les fragments d'ARN de transfert (tiARN) et les longs ARN non codants (lncARN)(Booy *et al.*, 2016 ; Ivanov *et al.*, 2014 ; Jayaraj *et al.*, 2012).

*rG4 aux télomères*

Les G4 d'ADN peuvent être formés dans les répétitions télomériques, mais les rG4 sont aussi importants pour l'activité de la télomérase. En effet, la télomérase est une ribonucléoprotéine avec une sous-unité protéique catalytique (hTERT) et une sous-unité ARN (hTR) qui permet de reconnaître son substrat et qui sert de matrice pour l'élongation des télomères. L'ARN hTR peut adopter un rG4 qui nuit au repliement de la structure secondaire fonctionnelle de la sous-unité et qui est reconnu par une hélicase (Gros *et al.*, 2008 ; Booy *et al.*, 2012). De plus, les régions des répétitions télomériques sont aussi transcrites en ARN ce qui forme les ARN non codants TERRA qui peuvent aussi adopter la formation rG4 et qui sont essentiels pour l'intégrité télomérique (Martadinata *et al.*, 2011).

### Transcription

Suivant l'effet de la formation des G4 d'ADN dans les télomères, la seconde fonction la plus étudiée des G4 est l'effet de leur présence dans les promoteurs (Huppert et Balasubramanian, 2007). Les G4 sont enrichis dans les régions proximales aux sites d'initiation de la transcription et globalement, la formation du G4 et sa stabilisation à l'aide de ligand spécifique inhibe la transcription du gène sur lequel il se retrouve (Cogoi et Xodo, 2006). De ce fait, les rG4 ont aussi été soupçonnés d'avoir un rôle dans la transcription. La formation d'une « *R-loop* » survient lorsque le brin d'ARN en cours de transcription s'apparie avec le brin d'ADN codant (antisens). Cela affecte la régulation de la transcription (Costantino et Koshland, 2015). Dans le cas des séquences rG4, ce sont les séries de G nouvellement transcrites qui peuvent interagir avec le brin non codant G-riche de l'ADN et ainsi former des hybrides G4 ADN-ARN qui sont assez stables et répressifs pour la transcription (Zheng *et al.*, 2013). Dans cette situation, seulement 2 séries de G sont nécessaires sur le brin d'ARN pour s'apparier avec 2 séries de G de l'ADN pour former les hybrides G4 intermoléculaires. Des études bio-informatiques du même groupe ont aussi démontré que les séquences susceptibles d'adopter ces hybrides sont aussi enrichies dans les régions proximales des sites d'initiation de la transcription. Ces sites sont conservés chez les animaux à sang chaud et sont donc suggérés pour être des éléments importants de la régulation de la transcription (Xiao *et al.*, 2013).

### Maturation des ARNm et régulation post-transcriptionnelle

La maturation des pré-ARNm survient de façon co-transcriptionnelle, par l'interaction du brin d'ARN avec une variété de protéines et de complexes liant l'ARN. Elle consiste à l'ajout d'une coiffe en 5' et d'une queue poly-A en 3', des ajouts essentiels pour protéger le transcrit de la dégradation par des exonucléases et assurer sa stabilité. Une autre étape majeure de la maturation est l'épissage du pré-ARNm, qui consiste à enlever les introns permettant d'obtenir la séquence de l'ARNm mature qui sera traduite en protéine (**Figure 1**). L'ensemble de la régulation post-transcriptionnelle qui suit survient directement auprès de l'ARNm suivant son interaction avec d'autres RBP permettant de réguler l'export du transcrit du noyau, ensuite sa localisation dans la cellule ou vers diverses vésicules et granules, moduler sa traduction et finalement sa dégradation. Des rG4 ont été identifiés comme

élément régulateur pour chacune de ces étapes de régulation co- et post-transcriptionnelle. Les effets des rG4 varient selon leur localisation dans le brin d'ARNm mature ou pré-messager, et selon leur localisation soit dans le 5'ou le 3'UTR, les introns, les exons ou dans la séquence codante.

**Ajout de la coiffe**

Une des premières étapes de la maturation des ARNm est l'ajout co-transcriptionnel de la coiffe de 7-méthylguanosine ($m^7G$) à l'extrémité 5' du transcrit via un pont 5',5'-triphosphate. Dans les cellules humaines, ce processus est régulé par deux enzymes ; la HCE (*human capping enzyme)* et une méthyltransférase (Galloway et Cowling, 2018). Il a été observé que plus la distance est courte entre un rG4 situé en 5'UTR et l'extrémité 5', plus l'effet inhibiteur du rG4 sur les niveaux d'expression du transcrit est grand. Ce qui laisse présager que le rG4 nuit à la synthèse de la coiffe $m^7G$, essentielle pour la stabilité du transcrit d'ARNm, ou encore que le rG4 nuit à la reconnaissance de la coiffe par les facteurs d'initiation de la traduction tels que eIF4E, la protéine liant la coiffe dans le complexe d'initiation de la traduction (Bugaut et Balasubramanian, 2012 ; Huppert *et al.*, 2008 ; Kumari *et al.*, 2008). Cependant, outre l'observation que la position du rG4 relative à l'extrémité 5' de l'ARNm corrèle avec les degrés de répression de l'expression de l'ARNm, l'impact direct de la liaison des facteurs d'initiation de la traduction au rG4 ou l'activité enzymatique d'ajout de la coiffe en présence de rG4 n'a pas été systématiquement étudié. Il est tenu pour acquis que pareillement à une structure secondaire très stable comme une tige-boucle située à l'extrémité 5'(Kozak, 1989), les rG4 peuvent être tout aussi répresseurs pour la reconnaissance de la coiffe et l'initiation de la traduction.

**Épissage**

L'épissage est l'étape de maturation du pré-ARNm qui consiste à enlever les segments d'introns non codants et de liguer les segments d'exons. L'épissage est un processus constitutif et plus de 90% des pré-ARNm subissent cette étape. L'épissage peut aussi être alternatif, c'est-à-dire que certains segments d'exons et d'introns sont alternativement conservés ou épissés résultant en différentes possibilités de transcrits matures à partir d'un seul pré-ARNm. Cela a pour fonction d'augmenter la diversité du transcriptome et des protéines résultantes, mais aussi de modifier la localisation ou les niveaux d'expression des

différents isoformes d'un transcrit (Wang *et al.*, 2008). La machinerie cellulaire responsable d'effectuer l'épissage est le spliceosome. Plusieurs facteurs d'épissage et des RBP reconnaissent des motifs sur le pré-ARNm indiquant les sites d'épissage et guident la machinerie afin de choisir les sites et d'effectuer le clivage des introns et la ligation des exons. La présence de structures secondaires, comme des tiges-boucles dans les pré-ARNm, particulièrement à proximité des sites d'épissage est reconnue pour affecter l'efficacité de ce processus. Des erreurs dans l'épissage ou encore des changements dans les ratios d'épissage alternatif peuvent entraîner des dérèglements cellulaires et des maladies (Soemedi *et al.*, 2017).

Les travaux initiaux de Maizel sur la densité des motifs G4 dans le génome ont montré que les G4 sont enrichis dans les premiers introns, suggérant un effet possible des structures secondaires rG4 sur l'épissage. Depuis, de nombreux travaux ont confirmé l'hypothèse que les rG4 affectent l'épissage. Le tout premier concerne l'épissage alternatif de l'ARN de la télomérase humaine hTERT qui est affecté par la présence de ligand stabilisant le rG4 (Gomez *et al.*, 2004). Par la suite, des rG4 ont été démontrés pour affecter l'épissage alternatif d'importants transcrits neuronaux comme BACE-1 impliqué dans la maladie d'Alzheimer et FMR1 dans la maladie du X-fragile (Didiot *et al.*, 2008 ; Fisette *et al.*, 2012). Les rG4 agissent aussi comme motifs régulateurs de l'épissage pour des transcrits essentiels dans la carcinogenèse en affectant l'épissage alternatif des transcrits du gardien du génome p53 (Marcel *et al.*, 2011 ; Perriaud *et al.*, 2014) ou des transcrits apoptotiques (Hai *et al.*, 2008). En stabilisant le rG4 avec des ligands spécifiques, il est même possible de stimuler ou non l'épissage, entre autres pour le transcrit Bcl-X dont les deux isoformes du transcrit, long et court ont respectivement des activités opposées anti- et pro-apoptotiques (Weldon *et al.*, 2017b). Dans la plupart des cas, les rG4 présents dans les introns sont reconnus par des RBP (**Tableau 1**), et ils peuvent agir autant à titre de stimulateur (*enhancer*) que de répresseur de l'épissage selon le transcrit.

**Terminaison de la transcription et polyadénylation**

L'ajout de la coiffe ainsi que l'épissage sont des processus de maturation des pré-ARNm qui surviennent de façon co-transcriptionnelle. La transcription du pré-ARNm prend fin lors de l'étape finale de la terminaison. Vers la fin du gène, la polymérase transcrit le site de polyadénylation (AAUAAA) suivi d'un site GU-riche situé légèrement en aval. Cela

constituera un signal de reconnaissance pour le complexe de terminaison de la transcription et une endonucléase viendra cliver le transcrit. L'extrémité 5' créée sera reconnue par l'exonucléase Xrn2 qui dégradera le transcrit rapidement en rejoignant la polymérase pour la faire « décrocher » selon le mécanisme de terminaison « torpille » (Watson *et al.*, 2009). Les rG4 stimuleraient la terminaison de la transcription encore une fois grâce à leur stabilité élevée qui ralentirait la progression de la polymérase favorisant ainsi la terminaison, le clivage et l'ajout de la queue poly-A. En utilisant l'outil de prédiction *quadparser,* il a été constaté que les PG4 étaient enrichis dans le 3'UTR des gènes possédant un second gène à moins de 1 Kb en 3' UTR, il a donc été proposé que ces rG4 permettraient de réduire le risque de « passer tout droit » (*read-through*) de la polymérase dans le gène suivant (Huppert *et al.*, 2008).

L'impact d'un rG4 est aussi connu pour la terminaison de la transcription de l'ARNm de l'*insulin growth factor II*, ainsi que pour conserver la terminaison correcte du 3'UTR de l'ARNm pré-messager de p53 lors de dommages à l'ADN (Christiansen *et al.*, 1994 ; Decorsiere *et al.*, 2011). Les rG4 sont aussi impliqués dans la terminaison de la transcription de l'ARN mitochondrial (Wanrooij *et al.*, 2010).

Les rG4 auraient aussi un rôle à jouer lors de la polyadénylation, l'étape de la maturation des ARNm qui survient dans le 3'UTR. Suite à la reconnaissance du site de polyadénylation (pAS) canonique AAUAAA et du clivage de l'ARNm en cours de transcription, l'enzyme poly-A polymérase vient synthétiser une série d'adénines d'environ 200 nt de long à la fin du transcrit. Cet ajout assure la stabilité du transcrit envers les exonucléases et sa reconnaissance par le système de contrôle de la qualité. Bien qu'il y ait un motif canonique de pAS, d'autres variations sont possibles. De plus, une séquence 3'UTR peut en contenir plusieurs. Cela entraîne donc une régulation pour sélectionner le site de polyadénylation et modifier la maturation de l'extrémité 3' du transcrit. Le choix de ce site peut entraîner des conséquences fonctionnelles, par exemple en utilisant un site plus près de la fin de la séquence codante, la région 3'UTR sera plus courte. Puisque la région 3'UTR peut être reconnue par des miARN pour réguler le transcrit soit en inhibant sa traduction ou en entraînant sa dégradation, la perte possible de ces sites de liaison miARNs en ayant un 3'UTR plus court peut rendre ces transcrits plus stables et plus exprimés. Les travaux de Jean-Denis Beaudoin dans le laboratoire du Pr Perreault ont permis de constater que le

positionnement relatif du rG4 par rapport aux sites de pAS en 3'UTR peut venir influencer le choix du site de polyadénylation pour deux transcrits, soit LRP5 et FXR1 (Beaudoin et Perreault, 2013).

**Localisation cellulaire**

Pour certains types cellulaires, la traduction de l'ARNm doit se faire à un moment et surtout à une localisation subcellulaire très précise. Particulièrement pour les neurones, l'ARNm transcrit au noyau doit être déplacé pour rejoindre l'autre extrémité de l'axone où il sera traduit à l'endroit requis pour la fonction de la protéine qu'il encode, par exemple à la synapse du neurone. Des rG4 situés dans le 3'UTR servent de motifs de localisation pour les ARNm PSD-95, CaMKIIα (Subramanian *et al.*, 2011), Anxa2 (Rihan *et al.*, 2017) et FMR1 (Schaeffer *et al.*, 2001), qui sont des protéines post-synaptiques importantes. Ces motifs rG4 sont reconnus par des RBP telles que FMRP (Phan *et al.*, 2011), hnRNP A2 (Sofola *et al.*, 2007), TDP-43 (Ishiguro *et al.*, 2016) et SMN (Rihan *et al.*, 2017) qui transportent ces ARNm. Les rG4 au sein de ces particules ribonucléoprotéiques (mRNP) permettent de réprimer la traduction lors du transport jusqu'à leur arrivée au site local de traduction.

**Traduction**

La séquence codante d'un ARNm (*coding sequence,* CDS) est composée de l'enchaînement des différents codons formés de trois nucléotides qui encodent l'ordre des acides aminés nécessaires pour la synthèse des peptides. L'étape de la traduction consiste donc à la lecture de ces codons et à la synthèse du lien peptidique entre chacun des acides aminés ajoutés en suivant la séquence, le tout effectué par la machinerie ribosomale à l'aide des ARN de transferts (ARNt) et des multiples facteurs de traduction. Ce processus cellulaire de synthèse est l'un des plus demandant énergétiquement pour une cellule et il est donc hautement régulé. L'étape majeure de régulation est l'initiation. C'est-à-dire la reconnaissance du codon d'initiation sur le transcrit d'ARNm et l'assemblage du ribosome complet. L'initiation s'effectue d'abord par la reconnaissance de la structure coiffe de l'ARNm par le complexe d'initiation de la traduction eIF4F formé des sous-unités eIF4E (protéine qui reconnaît la coiffe), eIF4A (hélicase) et eIF4G (protéine d'échafaudage). Suite au recrutement du complexe de pré-initiation (PIC) formé de la petite sous unité ribosomale 40S, du facteur eIF2, d'un GTP et de l'ARNt-Met (l'acide aminé de départ), le complexe se déplacera tout

au long de la région 5'UTR, une étape appelée « *scanning* », afin d'identifier le codon d'initiation AUG par appariement avec l'ARNt-Met (Sonenberg et Hinnebusch, 2009). Il est reconnu que la présence de structures secondaires stables dans le 5'UTR peut affecter l'efficacité du *scanning* et réprimer la traduction (Hinnebusch *et al.*, 2016 ; Leppek *et al.*, 2018). Les structures rG4 peuvent affecter la traduction selon des mécanismes qui varient selon leur positionnement dans le transcrit (**Figure 11**).

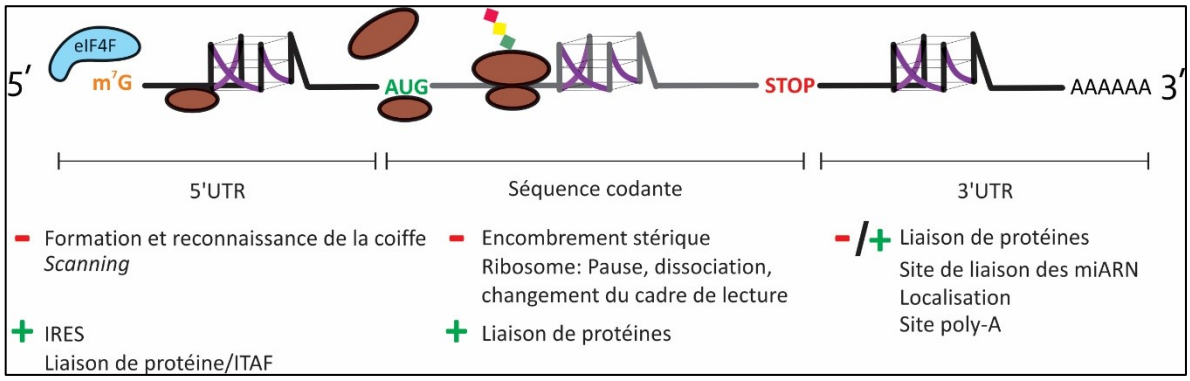



**Figure 11** – Effet des rG4 sur la traduction

Schéma résumé des effets répresseurs (signe « – » rouge) et activateurs (signe « + » vert) des rG4 dans la traduction selon leurs positions dans un transcrit d'ARNm. Les extrémités 5' et 3' du transcrit avec la coiffe $m^7G$ et la queue poly-A sont représentés. Le complexe de reconnaissance de la coiffe eIF4F est représenté en bleu, la sous-unité 40S du ribosome est représentée lors du *scanning* en 5'UTR (petit cercle brun). La sous-unité 60S (grand cercle brun) est assemblée lors de la reconnaissance du codon de départ AUG en vert pour former le ribosome actif 80S qui effectue la synthèse peptidique jusqu'au codon-stop indiqué en rouge.

Le **Tableau 2** présente en ordre chronologique de leur découverte les transcrits dont la traduction est affectée par la présence d'un rG4 selon son positionnement en 5'UTR, 3'UTR ou CDS. Les rG4 affectent la traduction directement en nuisant aux phases d'initiation et d'élongation ou indirectement en affectant la maturation de l'extrémité 3' ou la localisation cellulaire du transcrit tel que mentionné précédemment. La majorité des rG4 réprime la traduction, les transcrits indiqués en gras sont les seuls dont la présence du rG4 favorise leur expression.

**Tableau 2** Transcrits avec rG4 affectant leur traduction selon leur position

| Position | Transcrit | Référence |
|---|---|---|
| **5'UTR** | **FGF2** | (Bonnal *et al.*, 2003) |
| | NRAS | (Kumari *et al.*, 2007 ; Katsuda *et al.*, 2016) |
| | ZIC1 | (Arora *et al.*, 2008) |
| | ERS1 | (Balkwill *et al.*, 2009) |
| | MMP16 | (Morris et Basu, 2009) |
| | MAPK2, CHST2, PCGF2 | (Halder *et al.*, 2009) |
| | BCL2 | (Shahid *et al.*, 2010) |
| | AASDHPPT, BARHL1, EBAG9, FZD2, NCAM2, THRA | (Beaudoin et Perreault, 2010) |
| | TRF2 | (Gomez *et al.*, 2010) |
| | **VEGFA** | (Morris *et al.*, 2010 ; Cammas *et al.*, 2015 ; Bhattacharyya *et al.*, 2015) |
| | ADAM10 | (Lammich *et al.*, 2011) |
| | CCND3, YY1 | (Weng *et al.*, 2012) |
| | **TGFβ2** | (Agarwala *et al.*, 2013) |
| | AKTIP, CTSB, FOXE3 | (Agarwala *et al.*, 2014) |
| | USE1 | (Nieradka *et al.*, 2014) |
| | ARPC2, MMP16 | (von Hacht *et al.*, 2014). |
| | APC, HIRA, TOM1L2 | (Jodoin *et al.*, 2014 ; Jodoin et Perreault, 2018) |
| | H2AFγ, AKIRIN | (Rouleau *et al.*, 2015) |
| | ATR | (Kwok *et al.*, 2015) |
| | BNIP1, TEF | (Bolduc *et al.*, 2016) |
| | MST1R | (Ishiguro *et al.*, 2016) |
| | **NRF2** | (Lee *et al.*, 2017, p. 2) |
| | **SNCA** | (Koukouraki et Doxakis, 2016) |
| | HNF4A | (Guo et Lu, 2017, 2018) |
| | KRAS | (Miglietta *et al.*, 2017) |
| | SMNDC1 | (McAninch *et al.*, 2017) |
| | TAOK2, CXCL14 | (Zeraati *et al.*, 2017) |
| | SMXL4/5 | (Cho *et al.*, 2018) |
| | CHSY1 | (Yamaguchi *et al.*, 2018, p. 1) |
| | ARG2, AZIN1, **OAZ2**, ODC1, SMS | (Lightfoot *et al.*, 2018) |
| | BAG-1, CASP8AP2 | (Jodoin et Perreault, 2018) |
| **CDS** | FMR1 | (Schaeffer *et al.*, 2001) |
| | PRNP | (Olsthoorn, 2014) |
| | EBNA1 | (Murat *et al.*, 2014) |
| | MLL1, MLL4 | (Thandapani *et al.*, 2015) |
| | E4F1 | (Endoh et Sugimoto, 2016) |

| Position | Transcrit | Référence |
|---|---|---|
| **3'UTR** | CAMK2A, DLG4 | (Subramanian *et al.*, 2011) |
| | PIM1 | (Arora et Suess, 2011) |
| | TP53 | (Decorsiere *et al.*, 2011) |
| | SHANK1 | (Zhang *et al.*, 2011b) |
| | KISS1 | (Huijbregts *et al.*, 2012) |
| | FXR1, LRP5 | (Beaudoin et Perreault, 2013). |
| | APP | (Crenshaw *et al.*, 2015) |
| | AVPR1B, DOK1, KIF26A, PTPRU | (Bolduc *et al.*, 2016) |
| | ANXA2 | (Rihan *et al.*, 2017) |
| | ARG2, SMS, OAZ1, OAZ3 | (Lightfoot *et al.*, 2018) |

Les rG4 nuisent à l'initiation de la traduction lorsqu'ils sont situés en 5'UTR. Le rôle de répresseur de la traduction des rG4 situés en 5'UTR est l'un des plus observés. Cela ne pourrait correspondre qu'à la pointe de l'iceberg, puisque l'ensemble des méthodes de prédiction des rG4 s'accorde pour mentionner un enrichissement des PG4 dans les régions 5'UTR. La liste présentée au **Tableau 2** ne comporte qu'une fraction de ces prédictions. Cela démontre leur possible grande influence régulatrice dans la traduction.

Cependant, malgré ce grand nombre de rG4 identifiés comme répresseurs, le mécanisme expliquant cet effet est peu décrit. On statue souvent simplement que les rG4 sont des structures extrêmement stables qui bloquent les ribosomes lors du *scanning* et de l'élongation. Les toutes premières études sur les rG4 et la traduction semblent démontrer à l'aide de gènes rapporteurs que plus le rG4 est stable, plus la répression est grande. (Halder *et al.*, 2009 ; Wieland et Hartig, 2007). Tel que décrit dans la section « Ajout de la coiffe », Balasubramanian et son équipe ont observé que plus les rG4 étaient près de l'extrémité 5' et plus l'effet répressif était grand (Kumari *et al.*, 2008). L'effet répressif d'un même rG4 en 5'UTR sur la traduction varie légèrement en intensité, mais reste présent, et ce même s'il est testé dans différentes lignées cellulaires transformées ou non (Beaudoin et Perreault, 2010 ; Halder *et al.*, 2009). La stabilisation des rG4 situés en 5'UTR à l'aide de ligands réprime encore plus la traduction (Bugaut *et al.*, 2010). Donc, l'explication acceptée dans la littérature à ce jour est que les rG4 sont des structures très stables, ce qui ralentit et nuit au *scanning* du ribosome afin d'identifier le codon de départ.

Par contre, comment expliquer qu'un rG4 en 5'UTR est nécessaire pour la traduction de TGF-β2 (Agarwala *et al.*, 2013) ? Pour cet ARN, la présence du rG4 augmente sa traduction tandis qu'une mutation qui abolit le rG4 produit l'effet inverse. Une explication proposée est que les rG4 puissent faire partie de structure IRES (*internal ribosome entry site*).

Le *scanning* est une étape essentielle de la traduction dépendante de la reconnaissance de la coiffe dans le processus canonique d'initiation de la traduction qu'on appelle coiffe-dépendant (*cap*-dépendant). Il existe cependant un second mode d'initiation de la traduction qui est lui indépendant de la reconnaissance de la coiffe. Ce type d'initiation de la traduction, présent pour plusieurs virus, est aussi identifié pour des ARNm eucaryotes. Ce mécanisme requiert la présence d'une structure secondaire appelée IRES qui permet de recruter directement les facteurs d'initiation de la traduction sans la nécessité de posséder une coiffe ou de la reconnaître. Des rG4 ont été décrits comme des motifs essentiels dans des structures IRES notamment dans le 5'UTR de l'ARNm de FGF-2 (Bonnal *et al.*, 2003) et de VEGFA (Morris *et al.*, 2010). Cependant, l'effet du rG4 dans l'IRES de VEGFA est un sujet débattu puisqu'une étude a démontré le contraire, soit que sa stabilisation nuit à la traduction IRES-dépendante (Cammas *et al.*, 2015), alors qu'un autre groupe décrit le rG4 comme étant essentiel pour recruter la sous-unité 40S et initier la traduction cap-indépendante (Bhattacharyya *et al.*, 2015).

Lorsque les rG4 sont situés dans la région codante, leur effet sur l'élongation de la traduction est très clair. Ceux-ci ralentissent et nuisent à la progression du complexe d'élongation. La présence de cette structure stable peut entraîner des pauses, des décrochages ou encore entraîner un changement du cadre de lecture lors de la progression du ribosome (Endoh *et al.*, 2013c, 2013b ; Endoh et Sugimoto, 2013). Par contre, cela peut être bénéfique pour la maturation de la protéine qui est traduite. Ces pauses dans l'élongation permettent d'effectuer la protéolyse nécessaire à la maturation complète d'un récepteur à l'œstrogène (Endoh *et al.*, 2013a). Le niveau de répression de l'élongation est dépendant de la stabilité du rG4, mais aussi de son positionnement selon la périodicité de 3 nt des codons. Si le rG4 est situé en position 0, +3 ou +6 son effet sera moindre qu'en position +1 ou +2 (Endoh et Sugimoto, 2016). Somme toute, vu leur stabilité qui nuit à la progression ribosomale, cela explique pourquoi les motifs PG4 sont déplétés et peu prédits dans l'ensemble des régions codantes du génome.

Des rG4 localisés en 3'UTR ont aussi été décrits pour influencer la traduction. Outre les cas mentionnés précédemment où ces rG4 affectent la maturation de l'extrémité 3', la liaison de miARN et de RBP ou encore la localisation cellulaire du transcrit, le mécanisme de régulation de la traduction n'est pas élucidé. Puisque l'ARNm est circularisé lors de la traduction et donc que les extrémités du 5'UTR et du 3'UTR sont rapprochées afin de favoriser plusieurs rondes de traductions, il est plausible que des éléments rG4 en 3'UTR puissent affecter le recrutement de facteurs et la traduction ou qu'encore une fois leur stabilité entraîne un encombrement stérique défavorable.

**Dégradation des ARNm**

La vie d'un ARNm se termine par sa dégradation. Le transcrit est déadénylé et sa coiffe enlevée, puis il sera dégradé par des exonucléases. Il y a peu d'études concernant l'impact des rG4 sur la demi-vie ou la dégradation des ARNm. Il peut être tentant de croire que par leur grande stabilité structurale les rG4 pourraient prévenir la dégradation, mais il n'y a pas de preuves formelles à ce sujet. Par contre, des travaux récents ont décrit un nouveau processus de dégradation par clivage endonucléolytique des ARNm durant la traduction, lors de la sortie de l'ARN du canal du ribosome. Ce processus de dégradation a été intitulé *Ribothrypsis*. Selon les auteurs, les rG4 pourraient être un signal pour ce type de dégradation puisque les sites de clivage observés sont enrichis en séquence PG4 (Ibrahim *et al.*, 2018). De plus, les rG4 sont connus pour être des sites de pauses ou d'arrêt de la progression des ribosomes, un autre facteur important pour ce nouveau mécanisme de dégradation des ARNm.

## Rôles des G-quadruplexes dans le développement, les maladies et le cancer

Toutes les étapes de la régulation post-transcriptionelle de l'expression génique peuvent être affectées d'une façon ou d'une autre par la formation de rG4. Plusieurs de ces étapes sont connues pour être affectées dans différentes maladies, incluant des maladies neurodégénératives et des cancers. Tel que décrit dans les sections précédentes, plusieurs mécanismes moléculaires élucidés pour les rG4 impliquent des ARNm ou de RBPs dérégulés dans ces maladies comme la nucleolin, FMRP, des hnRNP et des hélicases (Armas et Calcaterra, 2018 ; Cammas et Millevoi, 2017 ; Maizels, 2015).

Les rG4 peuvent être affectés par une mutation dans leur séquence qui entraîne la perte de leur repliement. Il a été observé qu'une mutation ponctuelle ou un SNP (*single-nucleotide polymorphism)* était suffisante pour affecter le repliement d'un rG4 et l'expression d'un transcrit (Baral *et al.*, 2012 ; Beaudoin et Perreault, 2010), dont des transcrits importants dans la carcinogenèse (Zeraati *et al.*, 2017). À l'inverse, une mutation peut favoriser la formation d'un rG4. Ceci est le cas pour la multiplication des répétitions GGGGCC dans le gène C9ORF72 qui entraîne la formation de structures rG4 et d'agrégats d'ARN et de RBP associés aux maladies neurodégénératives de la sclérose latérale amyotrophique (ALS) et de la démence fronto-temporale (FTD) (Conlon *et al.*, 2016 ; Schludi et Edbauer, 2017). La protéine prion dont l'agrégation est associée à des troubles du système nerveux tel que la maladie de Creutzfeldt–Jakob est liée par des aptamères rG4 et son ARNm lui-même peut adopter une structure rG4 qui servirait à son auto-régulation (Cavaliere *et al.*, 2013 ; Olsthoorn, 2014).

Des essais en cellule avec l'anticorps spécifique au rG4 BG4 ont montré une augmentation du signal dans des cellules cancéreuses (Biffi *et al.*, 2014b). La présence des rG4 dans plusieurs transcrits d'oncogènes comme N-RAS et BCL2 affectent leur traduction, ce qui en fait des cibles thérapeutiques potentielles (Miglietta *et al.*, 2017). Des rG4 dans des lncARN ont aussi été montrés pour jouer un rôle dans la migration des cellules cancéreuses du colon (Matsumura *et al.*, 2017). Ceci est une courte énumération pour ne nommer que quelques exemples de l'implication des rG4 dans diverses maladies.

Les rG4 sont globalement beaucoup moins étudiés que leurs équivalents d'ADN. Pourtant, ils ont des rôles biologiques importants, qui peuvent être dérégulés lors de maladies neurodégénératives, de cancers, et durant les mécanismes d'infections virales et bactériennes. Heureusement, leurs particularités structurales permettent de les cibler avec des ligands. Il est donc primordial de bien comprendre leur réelle prévalence, les facteurs qui affectent leur formation et de comprendre plus en détail les mécanismes biologiques auxquels ils prennent part.

## Hypothèses et problématiques

Les études accumulées au commencement de mes études supérieures en janvier 2012, qui ont été abordées en introduction, suggéraient que la définition originale d'un motif G4 n'était pas adéquate pour les G4 situés dans les ARNm. La présence du motif G4 canonique seulement ne permet pas nécessairement son repliement (Beaudoin et Perreault, 2010). De plus, des résultats expérimentaux ont montré que les G4 pouvaient être des structures secondaires plus diversifiées que celles décrites par le motif canonique (Amrane *et al.*, 2012 ; Guédin *et al.*, 2010). Ensuite, des structures G4 ont été montrées comme moins statiques, plus malléables et dynamiques selon leurs séquences (Kumar *et al.*, 2008 ; Arora *et al.*, 2009 ; Saxena *et al.*, 2010) ou les conditions en solution (Khateb *et al.*, 2007 ; Bugaut *et al.*, 2012). Tout cela étant différent de ce qui était considéré auparavant dans les travaux initiaux (Huppert et Balasubramanian, 2005 ; Todd *et al.*, 2005 ; Huppert *et al.*, 2008 ; Halder *et al.*, 2009). À ce moment, les travaux de recherche sur les rG4 étaient très limités comparativement aux travaux sur les G4 d'ADN. Donc, en se basant uniquement sur la définition rigide des G4 canoniques, le nombre de rG4 serait donc à la fois sous-estimé et mal estimé. De ce fait, on néglige aussi leur importance biologique et les mécanismes d'action et d'interactions possibles avec d'autres éléments de la régulation post-transcriptionnelle des ARNm.

Puisque les rG4 sont enrichis dans les régions 5'UTR des ARNm et que plusieurs exemples d'ARNm dont la traduction est réprimée ou favorisée par la présence d'un rG4 sont connus, il semble que la formation de rG4 soit importante pour la régulation de la traduction. Il fallait donc développer des outils biochimiques et bio-informatiques fiables afin de mieux prédire et caractériser leur formation, permettant ainsi de valider les hypothèses concernant leur implication dans la traduction.

### Objectif #1

**Établir une méthode afin d'étudier le repliement G4 dans des séquences variées et des conditions plus représentatives du contexte biologique où ces structures se retrouvent.**

La majorité des études structurales sur les G4 ont été faites à partir de la séquence ADN G-riche des télomères, soit un motif simple et répété (TTAGGG) servant de base à la définition du motif canonique des G4. On n'a donc pas exploré les limites de la diversité

possibles des séquences formant des G4. De plus, ces évaluations structurales sont souvent limitées au motif minimal requis pour la formation du G4 soit les quatre séries de G et les boucles uniquement. La possibilité de compétitions pour la formation de structure secondaire canonique avec les nucléotides adjacents du G4 n'est donc pas considérée. Les méthodes classiques de spectroscopie nécessitent de grandes concentrations (µM) d'ARN afin d'obtenir un signal interprétable et sont donc réalisées dans des conditions qui ne représentent pas les conditions biologiques et qui favorisent la formation des structures intermoléculaires plutôt qu'intramoléculaires. Il est donc difficile d'évaluer rapidement le repliement ou non d'un G4 dans un contexte similaire au contexte biologique. Puisque les séquences formant les G4 d'ARN semblent aussi être plus variées en termes de nombre de séries de G impliquées, du nombre de tétrades empilées et des tailles possibles des boucles, et que ces particularités de leur structure affectent leur stabilité, leur reconnaissance par des facteurs protéiques et leurs fonctions biologiques, une méthode permettant de caractériser en détail et rapidement ces diverses spécificités est nécessaire. Cela permettra de tester les limites de repliement rG4 en évaluant la structure secondaire de motifs PG4 possédant des caractéristiques plus variées comme des boucles plus longues ou un contexte nucléotidique flanquant plus étendu.

**Objectif #2**

**Développer un meilleur outil de prédiction des G4 d'ARN basé sur des facteurs pouvant affecter leur repliement autres que la simple présence du motif canonique.**

Plusieurs séquences d'ARN respectant l'algorithme classique de prédiction de PG4 : $G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$, n'adoptent pas de structure rG4 lorsqu'elles sont évaluées expérimentalement, ce qui représente plusieurs faux positifs. Cela signifie que la séquence primaire ne constitue pas l'unique facteur permettant le repliement quadruplex. Afin d'améliorer les prédictions et d'identifier correctement la prévalence des rG4 dans le transcriptome, l'objectif est donc d'identifier ces facteurs supplémentaires et de trouver une façon rapide de pouvoir les prendre en considération dans la prédiction.

À l'inverse, plusieurs séquences divergeant de l'algorithme sont, elles, démontrées pour adopter la structure rG4. Cela représente des faux négatifs selon les méthodes actuelles de prédiction. Tel que mentionné dans le premier objectif, on ignore quelles sont les limites

des séquences pouvant former des rG4 ainsi que le contexte permettant l'adoption de cette structure. L'outil de prédiction devra donc être « permissif » afin de pouvoir identifier ces rG4 « non canoniques ».

**Objectif #3**

**Déterminer si les G4 d'ARN en 5'UTR sont enrichis dans des voies biologiques particulières et par quels mécanismes ils affectent l'expression des ARNm sur lesquels ils se retrouvent.**

La littérature scientifique actuelle indique majoritairement que les rG4 situés dans les 5'UTR identifiés à ce jour sont des répresseurs de la traduction. Ceci est basé sur l'assomption logique que leur stabilité élevée nuit au *scanning* du ribosome et à l'initiation de la traduction. Pourtant, certains exemples de rG4 en 5'UTR favorisant la traduction sont connus. Le mécanisme d'action n'est donc peut-être pas si évident. On ignore si les ARNm connus pour être traductionnellement réprimés par des rG4 en 5'UTR sont des exemples particuliers (des exceptions ou des artefacts), ou s'ils sont synonymes d'un mécanisme de régulation plus complexe, basé sur la reconnaissance d'un motif rG4 particulier par des chaperonnes ou des hélicases pour coréguler des ARNm associés à une même voie biologique.

De plus, la traduction étant régulée principalement au niveau de l'initiation, plusieurs autres motifs : éléments *cis*-régulateurs, structures secondaires et facteurs de régulation protéiques sont connus pour jouer un rôle important dans cette régulation. Dans certaines conditions de stress ou dans des cellules cancéreuses, par exemple, plusieurs mécanismes de régulation non canoniques de la traduction sont connus. L'interaction entre les motifs rG4 et ces autres motifs de régulation est un aspect non exploré du champ de recherche sur les G4 d'ARN.

# ARTICLE 1– IN-LINE PROBING OF RNA G-QUADRUPLEXES

**Auteurs de l'article** : Beaudoin, Jean-Denis*, Jodoin, Rachel* et Perreault, Jean-Pierre
* Co-premiers auteurs

**Statut de l'article** : Publié dans Methods (2013), vol. 64, p. 79–87

**Avant-propos** : Rachel Jodoin a réalisé les expériences et les analyses présentées dans l'article. Jean-Denis a initialement établi la méthode *in-line* dans le laboratoire. Les figures ont été réalisées par Rachel Jodoin. L'article a été rédigé par Jean-Denis Beaudoin, Rachel Jodoin et Jean-Pierre Perreault.

**Résumé**

Malgré le fait que la majorité des études initiales sur les G-quadruplexes aient été effectuées avec des molécules d'ADN, il existe présentement un intérêt grandissant envers l'étude de ces structures, qui ont un fort potentiel d'agir comme élément régulateur de l'expression génique, dans les molécules d'ARN. En effet, les G-quadruplexes retrouvés dans les régions 5' non-traduites des ARNm sont répandus dans le transcriptome humain et agissent en tant que répresseurs généraux de la traduction. En plus de leur effet sur la régulation de la traduction, plusieurs autres étapes de la maturation des ARNm telles que l'épissage, la polyadénylation et la localisation sont influencées par la présence de G-quadruplexes d'ARN. Des approches bio-informatiques ont permis l'identification de milliers de séquences possibles de G-quadruplex d'ARN dans le transcriptome humain. Clairement, il y a un besoin pour le développement de méthodologies et de techniques rapides, simples et informatives afin de déterminer *in vitro* la capacité de repliement en G-quadruplex de ces séquences, ainsi que celle d'autres éléments régulateurs potentiels. Ce rapport décrit une méthodologie intégrée afin de mesurer la formation de G-quadruplex d'ARN qui combine l'utilisation d'algorithme bio-informatique, la prédiction de structure secondaire, la cartographie *in-line* avec analyse semi-quantitative, ainsi que l'utilisation de logiciels de représentation structurale. La puissance de cette approche est illustrée, étape par étape, suivant la

détermination de la structure secondaire adoptée par la séquence G-quadruplex potentielle retrouvée dans le 5'UTR de l'ARNm du gène *cAMP responsive element modulator* (CREM). Les résultats démontrent sans ambiguïté que la séquence de CREM adopte une structure G-quadruplex en présence de concentration physiologique d'ions potassium. Cette méthode de cartographie *in-line* est facile d'utilisation, robuste, reproductible et informative pour l'étude de la formation des G-quadruplex d'ARN.

**Abstract**

Although the majority of the initial G-quadruplex studies were performed on DNA molecules, there currently exists a rapidly growing interest in the investigation of those formed in RNA molecules that possess high potential of acting as gene expression regulatory elements. Indeed, G-quadruplexes found in the 5'-untranslated regions of mRNAs have been reported to be widespread within the human transcriptome and to act as general translational repressors. In addition to translation regulation, several other mRNA maturation steps and events, including mRNA splicing, polyadenylation and localization, have been shown to be influenced by the presence of these RNA G-quadruplexes. Bioinformatic approaches have identified thousands of potential RNA G-quadruplex sequences in the human transcriptome. Clearly there is a need for the development of rapid, simple and informative techniques and methodologies with which the ability of these sequences, and of any potential new regulatory elements, to fold into G-quadruplexes *in vitro* can be examined. This report describes an integrated methodology for monitoring RNA G-quadruplexes formation that combines bioinformatic algorithms, secondary structure prediction, in-line probing with semi-quantification analysis and structural representation software. The power of this approach is illustrated, step-by-step, with the determination of the structure adopted by a potential G-quadruplex sequence found in the 5'-untranslated region of the cAMP responsive element modulator (CREM) mRNA. The results unambiguously show that the CREM sequence folds into a G-quadruplex structure in the presence of a physiological concentration of potassium ions. This in-line probing-based method is easy to use, robust, reproducible and informative in the study of RNA G-quadruplex formation.

## 1. INTRODUCTION

G-rich DNA and RNA molecules can form a non-canonical tetrahelical structure called a G-quadruplex (Burge *et al.*, 2006 ; Millevoi *et al.*, 2012). The primary building block of this structure is named a G-quartet and is composed of four coplanar guanines that form Hoogsteen base pairs involving a total of eight hydrogen bonds (Gellert *et al.*, 1962). These quartets are stabilized by a central counterion, typically potassium, and stack one on top of the other forming a very stable tetrahelical G-quadruplex structure (Huppert, 2008a). There is evidence that this structure forms *in cellulo* and that it is frequently found, at both the DNA and RNA levels, in cellular regulatory sequences such as promoters, telomeres and 5'-UTRs (Lipps et Rhodes, 2009). Many G-quadruplexes have been found to be associated with cell disorders, and, therefore, they constitute good potential therapeutic targets (Collie et Parkinson, 2011).

While most of the early G-quadruplex studies were performed on DNA molecules, more recently a rapidly growing interest has emerged in investigating those formed in RNAs. Moreover, recent research has revealed that the size of the cellular transcriptome is considerably larger than previously thought, with results showing that over 90% of the human genome is actively transcribed (ENCODE Project Consortium, 2007). In this new context, the importance of post-transcriptional regulation events is now appreciated more than ever. RNA G-quadruplexes are widely found in the cell and have been shown to act as efficient post-transcriptional regulatory elements that are involved in various biological mechanisms. These include: mRNA splicing, polyadenylation, translation and localization (Beaudoin et Perreault, 2010 ; Bugaut et Balasubramanian, 2012 ; Decorsiere *et al.*, 2011 ; Marcel *et al.*, 2011 ; Subramanian *et al.*, 2011). Several thousand potential RNA G-quadruplex sequences have been identified within the human transcriptome (Beaudoin et Perreault, 2010 ; Huppert *et al.*, 2008 ; Kikin *et al.*, 2008). In order to test the ability of this plethora of RNA sequences to fold into G-quadruplexes, the development of a simple, reliable and reproducible technique is required.

X-ray crystallography and NMR experiments have successfully produced high-resolution structures of many G-quadruplexes which provide a significant amount of information about these structures (Neidle et Parkinson, 2008). However, these techniques are time consuming and describe only one of the structures that can be formed by a given

sequence, a structure which does not necessarily correspond to the most abundant one. Clearly, quicker experiments studying the entire population of structures formed in solution must be considered. Circular dichroism (CD) is extensively used to monitor G-quadruplex formation (Randazzo *et al.*, 2012). Of particular importance, CD is able to distinguish parallel structures from antiparallel ones. Depending on its topology, the G-quadruplex structure exhibits characteristic spectral features in CD. Typically, a spectrum exhibiting a positive peak at a wavelength around 264 nm and a negative one around 240 nm is indicative of a parallel structure, whereas a spectrum showing positive peak at 295 nm and a negative one around 260 nm indicates the presence of an antiparallel structure. Since other nucleic acid structures can produce a positive peak around 260 nm, it is important to compare spectra recorded under conditions unfavorable for G-quadruplex formation (e.g. either in the absence of salt, or in the presence of $Li^+$ which acts inefficiently as the G-quadruplex counterion) with others recorded under favorable conditions (e.g. in the presence of either $Na^+$ or $K^+$). A transition to a characteristic G-quadruplex spectrum has to be observed between these conditions in order to suggest the formation of this particular structure. Alternatively, thermal denaturation is also commonly used to study G-quadruplex formation. It corresponds to a melting transition caused by an increase in temperature that can be monitored by either CD (e.g. at 264 nm for the parallel structure), or by the absorbance of UV light at 295 nm (Mergny et Lacroix, 2009). These values allow the determination of the melting temperature (Tm, temperature at which half of the structures are denatured). For sequences able to fold into G-quadruplexes, the calculated Tm are typically higher under favorable conditions (e.g. in the presence of either $Na^+$ or $K^+$), reflecting the prominent stability of these structures, as compare to those obtained under unfavorable conditions (e.g. either in the absence of salt or in the presence of $Li^+$). One of the limits of both the CD and thermal denaturation techniques is that they require relatively high concentrations of either DNA or RNA (i.e. in the low micromolar range). At these concentrations, both intra- and intermolecular G-quadruplex structures can easily be formed. As a result, neither of these two techniques can distinguish these two G-quadruplex topologies.

In-line probing is one of the simplest RNA structure chemical mapping techniques available (Regulski et Breaker, 2008). This technique is based on the tendency of RNA to be differentially hydrolyzed according to its structure (Soukup et Breaker, 1999). The

phosphodiester bonds of the RNA backbone are susceptible to slow, non-enzymatic cleavage through the “in-line” nucleophilic attack of the 2’-oxygen of the adjacent phosphorus group. This attack occurs when the 2’-oxygen, the phosphorus and the adjacent 5’oxygen adopt an “in-line” conformation that allows the 2’-oxygen to act as a nucleophile and to efficiently cleave the RNA linkage. Following this logic, the relative rate of spontaneous cleavage is directly related to the surrounding structural features of each RNA linkage. The flexible nucleotides, that is to say those found in single-stranded regions and at the periphery of the RNA structure, are free to adopt various conformations, including the “in-line” geometry, and, consequently, are more susceptible to cleavage. This approach has been extensively used to study both riboswitch secondary structures and the conformational changes that occur upon ligand binding (Regulski et Breaker, 2008).

A recent study, demonstrated the potential of in-line probing in monitoring the formation of intramolecular RNA G-quadruplex structures (Beaudoin et Perreault, 2010). It appears to be a very simple, reproducible and informative technique with which to study this motif. Since intramolecular RNA G-quadruplexes are forced to fold into parallel topologies due to their 2’-hydroxyl, C3’-*endo* sugar pucker and *anti* glycosidic bond geometry, they are typically composed of three external loops connecting the guanosine tracts (Burge *et al.*, 2006). The nucleotides located in these loops characteristically become highly flexible and are thus more susceptible to spontaneous cleavage upon G-quadruplex formation. This article describes a detailed integrated approach to the study of RNA G-quadruplex formation based on in-line probing analysis. This methodology takes advantage of bioinformatic algorithms for the identification of potential G-quadruplex (PG4) structures, a secondary structure prediction program, in-line probing and both quantification and structural representation software. In order to illustrate the procedure, the PG4 sequence found in the 5’UTR of the cAMP (cyclic adenosine monophosphate) responsive element modulator (CREM) mRNA was analyzed. This gene encodes a bZIP transcription factor that binds to the cAMP responsive element found in many promoters (Lamas *et al.*, 1996).

# 2. MATERIAL AND METHODS

## 2.1. Designing PG4s

Initially, *in vitro* PG4 versions are designed according to a potential G-quadruplex (PG4) sequence identified by a typical bioinformatic approach using the algorithm $G_x$-$N_{1–7}$-$G_x$-$N_{1–7}$-$G_x$-$N_{1–7}$-Gx, where x ≥ 3 and N is any nucleotide (A, C, G or U) (Wong *et al.*, 2010). Extra sequences of about 15 nucleotides are added to both the 5' and the 3' sides of the PG4 motif (see **Figure 12**). The nature of these sequences are identical to those found in the genomic regions flanking the PG4 in question. The main purpose of using extended *in vitro* PG4 versions is to render the analysis more biologically relevant. A previous study reported evidence that both the primary and secondary RNA structure contexts in the vicinity of the G-quadruplex structure were critical to RNA G-quadruplex formation both *in vitro* and *in cellulo* (Beaudoin et Perreault, 2010).

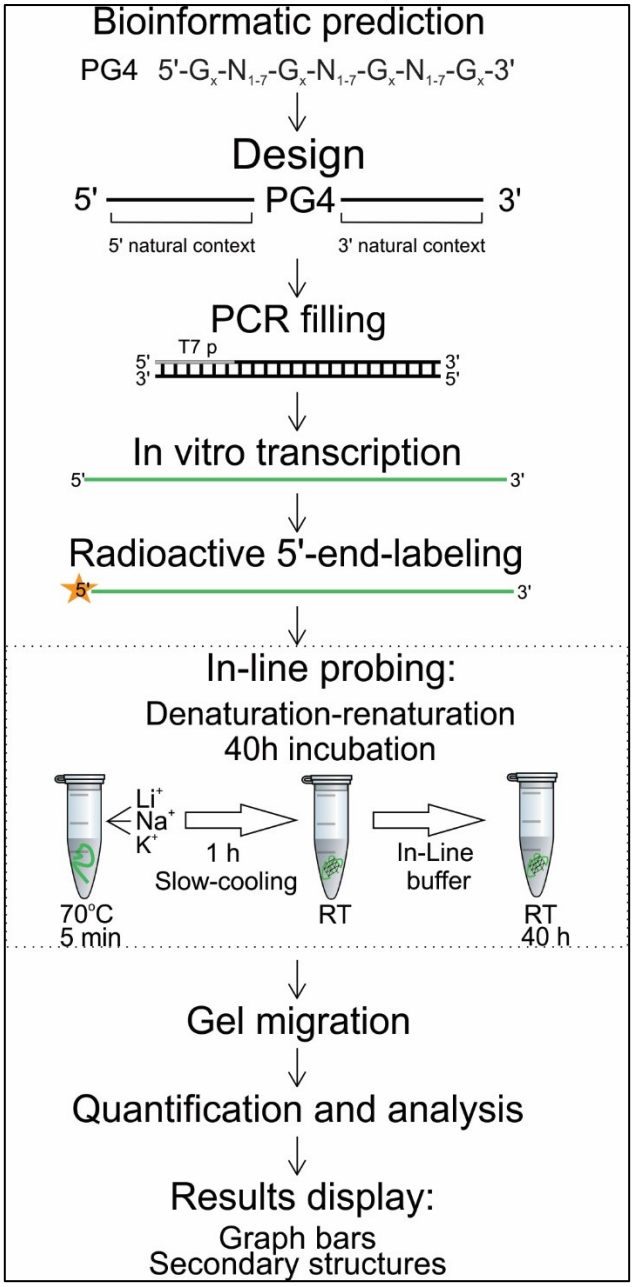



**Figure 12** – Organigram of the integrative approach to the study of RNA G-quadruplex formation using in-line probing.

In addition to the wild-type (wt) PG4 version, a mutated version in which some key guanines are substituted for by adenines (G/A-mut) must also be synthesized. It is important to disrupt most of the guanosine tracts of the PG4, as well as to consider the presence of any supplemental guanosine tracts located in either the 5' or 3' side (for example see **Figure 13A**). The G/A-mutant is a good negative control for G-quadruplex formation as it possesses only minor changes in its RNA sequence as compared to that of the wild type.

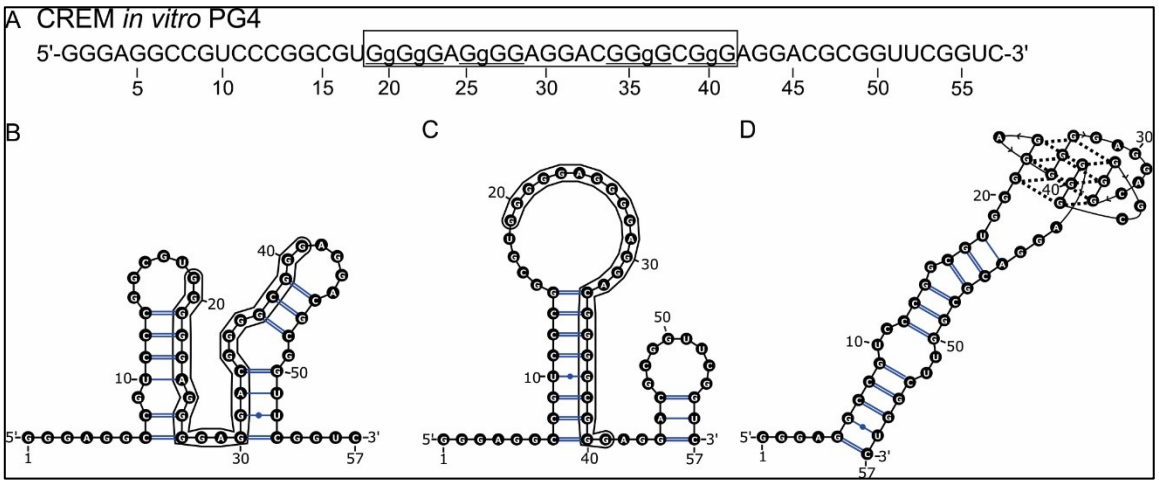



**Figure 13** – CREM PG4 sequence and its predicted secondary structures.

(A) Nucleotide sequence of the characterized CREM wt transcript. The boxed sequence denotes the predicted PG4. The four G-tracts are underlined. The lowercase guanines (g) correspond to those substituted for by adenines in the G/A-mutant. (B–C) The two secondary structures predicted by the RNAstructure software (version 5.4) for the wt CREM. (D) Possible secondary structures of the additional 5' and 3' regions of the PG4, predicted using the RNA structure software and combined with the representation of the unimolecular G-quadruplex structure predicted using the algorithm seeking potential PG4 sequences.

### 2.2. Secondary structure prediction

RNA secondary structure prediction software can be useful form both comparing and analyzing the in-line probing results. The predicted secondary structures of the candidate's *in vitro* PG4 version are retrieved using the RNAstructure software version 5.4 with the default settings (Reuter et Mathews, 2010). For the CREM PG4 wt version, the two predicted secondary structures with the lower energy values were then manually transposed into dot-and-bracket notations and pictured using the VARNA visualization applet (**Figure 13B and C**). A second secondary structure prediction was performed in order to determine the predicted structures adopted by the added 5' and 3' flanking sequences (i.e. by using an input sequence in which the potential G-quadruplex was substituted for by multiple adenines, thereby forcing it to form a large loop). The result of this second prediction is shown in **Figure 13D**, except that the large adenine loop was replaced for by a representation of the unimolecular parallel G-quadruplex structure predicted by the algorithm seeking G-quadruplex structures (i.e. by taking into account both the length of the G-tracts and the compositions of loops 1, 2 and 3). These various representations of the predicted secondary

structures are based on either strictly Watson–Crick base pairs (**Figure 13B and C**), or on structures that also include the formation of a G-quadruplex structure (**Figure 13D**), and can be used as an aid in analyzing the results obtained from the *in vitro* experiments.

**2.3. RNA synthesis**

After the selection of a candidate, and the design and analysis of the *in vitro* PG4 extended version, the next step is the production of the proper RNA molecules (**Figure 12).** Transcripts are synthesized by *in vitro* transcription using T7 RNA polymerase. First, two partially complementary DNA oligonucleotides (2 µM each, Invitrogen) are annealed and double-stranded DNA is obtained by filling the gaps using purified *Pfu* DNA polymerase in the presence of 5% dimethyl sulfoxide (DMSO, FisherScientific). One oligonucleotide corresponds to the reverse complementary sequence of the *in vitro* PG4 version with the addition of the 17 nucleotide (nt) reverse sequence of the T7 RNA polymerase promoter at the 3'end, while the other corresponds to the 17 nt sequence of the T7 RNA polymerase promoter extended by two or more guanines at the 3' end. In order to obtain good transcription efficiency, the T7 RNA polymerase requires the presence of a minimum of two guanines immediately 5' of the transcript. If these guanosines are not present within the natural PG4 sequence, the minimal number of guanines must be added in order to fulfill this requirement. This point should be taken into consideration in the designing of the *in vitro* PG4 versions. The DNA duplex containing the T7 RNA polymerase promoter sequence followed by the PG4 sequence is then ethanol-precipitated, ethanol-washed and dissolved in ultrapure water (Barnstead Nanopure). Run-off *in vitro* transcription reactions are then performed in a final volume of 100 µL using purified T7 RNA polymerase (10 µg) in the presence of RNase OUT (20 U, Invitrogen), pyrophosphatase (0.01 U, Roche Diagnostics) and 5 mM NTP (Sigma-Aldrich) in a buffer containing 80 mM HEPES-KOH, pH 7.5, 24 mM $MgCl_2$, 40 mM DTT (Fisher Scientific) and 2 mM spermidine (BioShop). The reactions are incubated for 2 h at 37°C, and are then treated with DNase RQ1 (Promega) at 37°C for 15 min (**Figure 12**). The resulting RNAs are then purified by phenol-chloroform extraction followed by ethanol precipitation. The RNA products are fractionated by denaturing (8 M urea) 10% polyacrylamide gel electrophoresis (PAGE ; 19:1 ratio acrylamide to bisacrylamide) using 45 mM Tris–borate pH 7.5 and 1 mM EDTA (BioShop) solution as running buffer. The transcripts are detected by UV shadowing, and the gel slices

containing those corresponding the correct sizes of the *in vitro* PG4s are excised. These slices are then incubated overnight at 4°C on a rotating wheel in a buffer containing 1 mM EDTA (Bio-Shop), 0.1% SDS (BioShop) and 0.5 M ammonium acetate (FisherScientific). The eluted RNAs are then ethanol-precipitated, dried, dissolved in ultrapure water and analyzed by spectrometry at 260 nm in order to determine their concentrations.

**2.4. Radioactive 5'-end labeling**

The next step is to radioactively label the RNA transcripts (**Figure 12**). In order to produce 5'-end-labeled RNA molecules, 50 pmol of purified transcripts are dephosphorylated at 37°C for 30 min in the presence of 1 U of antartic phosphatase (New England BioLabs) in a final reaction volume of 10 µL containing 50 mM Bis-propane (pH 6.0), 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$ and RNase OUT (20 U, Invitrogen). The enzyme is then inactivated by incubating for 7 min at 65°C. The dephosphorylated RNAs (10 pmol) are then 5'-end-radiolabeled using 3 U of T4 polynucleotide kinase (Promega) for 1 h at 37°C in the presence of 3.2 pmol of [$\gamma$-$^{32}$P]ATP (6000 Ci/mmol ; New England Nuclear). The reactions are stopped by the addition of 10 µL formamide dye buffer (95% formamide, 10 mM EDTA, 0.025% bromophenol blue and 0.025% xylene cyanol). Finally, the samples are purified by 10% polyacrylamide 8 M urea denaturing gel electrophoresis. The bands corresponding to the 5'-end-labeled RNAs are detected by autoradiography, and the gel slices containing those of the correct sizes are excised and recovered as described in the RNA synthesis (Section 2.3). The eluted and precipitated 5'-end-labeled transcripts are then dissolved in 30 µL ultrapure water, and the final radioactivity is calculated using a scintillation counter (Bioscan QC-2000).

**2.5. In-line probing experiment**

Prior to performing the in-line probing experiment, all 5'-end-labeled RNAs (both wt and G/A-mut PG4 versions) are heat-denatured and then allowed to slowly renature (**Figure 12**). More specifically, trace amounts of 5'-end-labeled transcripts (50000 cpm, <1 nM) are heated at 70°C for 5 min, and are then slow-cooled to room temperature over 1 h in a buffer containing 20 mM lithium cacodylate pH 7.5 and 100 mM of LiCl, NaCl or KCl, depending on the conditions tested, in a final volume of 10 µL. Following the initial slow-cooling step, the volume of each sample is adjusted to 100 µL such that the final concentrations are 20 mM lithium cacodylate pH 8.5, 20 mM $MgCl_2$ and 100 mM of LiCl, NaCl or KCl. The reactions

are then incubated for 40 h at room temperature, at which point the samples are ethanol-precipitated in presence of glycogen, ethanol-washed and dissolved in 10 µL ice-cold formamide loading buffer (95% formamide and 10 mM EDTA, 0.025% xylene cyanol).

Two ladders should be used for this kind of in-line probing experiment, an alkaline hydrolysis (permits the mapping of each nucleotide of the sequence) and an RNase T1 digestion of the transcripts (permits the mapping of the guanines). For the alkaline hydrolysis ladder, 50 000 cpm of the 5'-end-labeled wt transcripts (<1 nM) are dissolved in 5 µL of water, 1 µL of 1 N NaOH is added and the reaction is incubated for 1 min at room temperature prior to being quenched by the addition of 3 µL of 1 M Tris–HCl (pH 7.5). The RNA molecules are then ethanol-precipitated, and the RNA pellet dissolved in 10 µL formamide dye loading buffer (95% formamide, 10 mM EDTA and 0.025% xylene cyanol). For the RNase T1 ladder, 50 000 cpm of 5'-end-labeled wt transcript (<1 nM) are dissolved in 9 µL of buffer containing 20 mM Tris–HCl (pH 7.5), 10 mM $MgCl_2$ and 100 mM LiCl. The reaction mixture is incubated for 2 min at 37°C in the presence of 0.6 U of RNase T1 (Roche Diagnostic). The reaction is then quenched by the addition of 20 µL of formamide loading buffer (95% formamide, 10 mM EDTA and 0.025% xylene cyanol). All of the samples and ladders are then transferred into new microcentrifuge tubes, and the radioactive content of the in-line probing samples and both ladders are then quantified using a scintillation counter (Bioscan QC-2000). Equal amounts, in terms of cpm, of all samples ($Li^+$, $Na^+$, $K^+$), and approximately two-thirds of this amount of the ladders, are then fractionated on 10% polyacrylamide 8 M urea denaturing gels. The resulting gels are subsequently dried and visualized by exposure to phosphorscreen (GE Healthcare) using a Typhoon Trio instrument (GE Healthcare) (see **Figure 14** for an example).

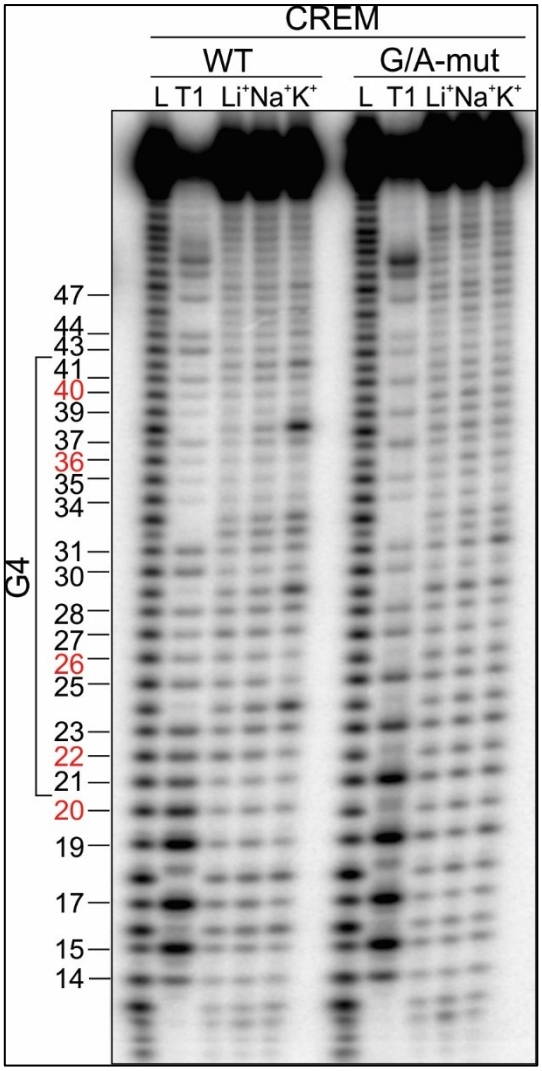



**Figure 14** – In-line probing results

Autoradiogram of a 10% denaturing (8 M urea) polyacrylamide gel of the in-line probing of both the 5'-labelled CREM wr and the G/A-mutant PG4 versions performed in the presence of 100 mM of either LiCL, NaCl or KCl. The L and T1 lane indicate the alkaline hydrolysis and ribonuclease T1 mapping lanes, respectively. The positions of the guanines are indicated at the left. The numbers in red represent guanines converted to adenines in the G/A-mutant version. The bracket at the left indicates the nucleotides involved in the formation of the G-quadruplex in the wt version.

### 2.6. Data analysis

Several types of data can be extracted from in-line probing gels (**Figure 12**). Initially, each gel is analyzed using the Semi-Automated Footprinting Analysis (SAFA) software (Laederach *et al.*, 2008). The RNase T1 ladder lane is used as the "anchor" line, using the guanines as cleavage sites for the sequence references in SAFA. The intensity of each band in each condition is determined and is exported in a text format file. This file can be opened

with the Excel program in order to produce an easily usable table. First, the intensity of the bands under the $Li^+$ conditions are used to examine the secondary structure adopted under conditions unfavorable to G-quadruplex formation. The intensities are normalized with a method commonly used for SHAPE structure probing (Low et Weeks, 2010). Briefly, the intensities of the bands having the next highest 10% intensities after the highest 2%, which corresponds to positions that are highly prone to cleavage, are averaged and each band's intensity is divided by this number, giving a ratio ranging between 0 and ~1. Low ratios correspond to constrained positions (i.e. mainly base-paired positions), while higher ratios indicate positions of greater flexibility such as single-stranded nucleotides. Normalization is performed using the data from 3 independent experiments, and the results are presented as a color map on the best predicted secondary structures (see **Figure 15A and B** for an example of the results obtained with the CREM candidate; blue indicates constrained nucleotides and red highly flexible ones).

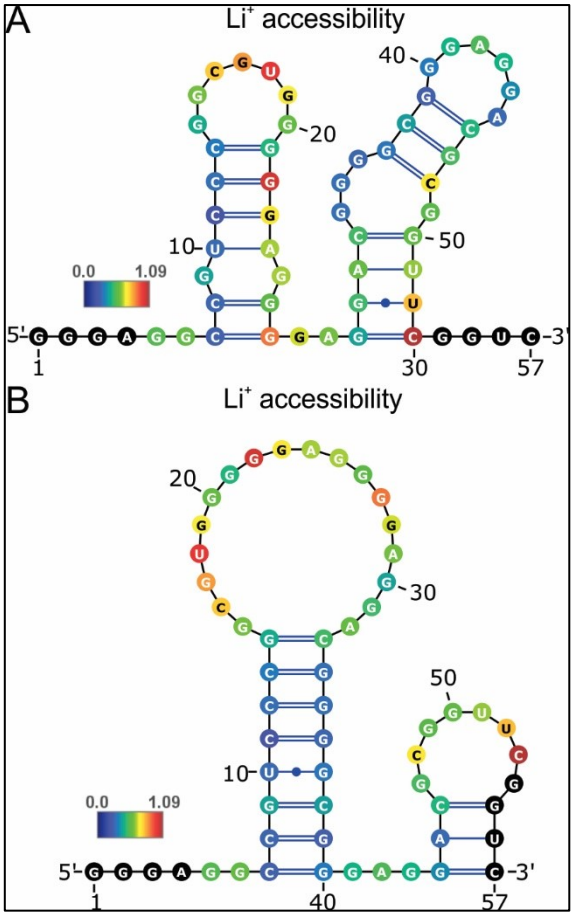



**Figure 15** – Nucleotide accessibility in the presence of $Li^+$.

(A) The first and (B) the second predicted secondary structures of the CREM wt. The color map illustrates each nucleotides' accessibility based on the normalization of the intensity of the band corresponding to each nucleotide, and is obtained by dividing by the average intensity of the 10% most intense bands. Ratios of ~0 (blue) show constrained regions, while ratios of ~1 (red) show flexible regions. The nucleotides in blacks are those for which no significant ratio could be calculated because their representative bands either migrated off of the gel, or were not sufficiently resolved.

Secondly, both the similarities and the discrepancies of the RNA structures under conditions both favorable and unfavorable for the formation of G-quadruplex structures are determined and examined. In order to do this, the raw intensity of each band from the lanes representing favorable G-quadruplex conditions (i.e. in the presence of either $Na^+$ or $K^+$) is divided by the intensity of the corresponding band from the $Li^+$ lane (i.e. the unfavorable condition). The in-line probing experiments are performed in triplicate (Materials and Methods point 2.5), and are then analyzed for each sequence (i.e. both the wt and the G/A-mut PG4s). The averages and standard deviations are calculated for the $Na^+/Li^+$ and $K^+/Li^+$

ratios for each nucleotide. These values are then used to generate bar graphs (with the intensity ratios of $Na^+/Li^+$ or $K^+/Li^+$ on the *y*-axis and the sequence on the *x*-axis) which permit an easier analysis of the data (see **Figure 16A** for the wt and **Figure 16B** for the G/A-mut sequences). Another way to represent the probing data is to show them directly on the predicted secondary structure. The same set of values can be used to create a color code in which the color of each nucleotide represents its cleavage susceptibility under conditions favorable for the formation of G-quadruplex structures relatively to that found under conditions that are unfavorable ($K^+/Li^+$) (**Figure 16C and D**).

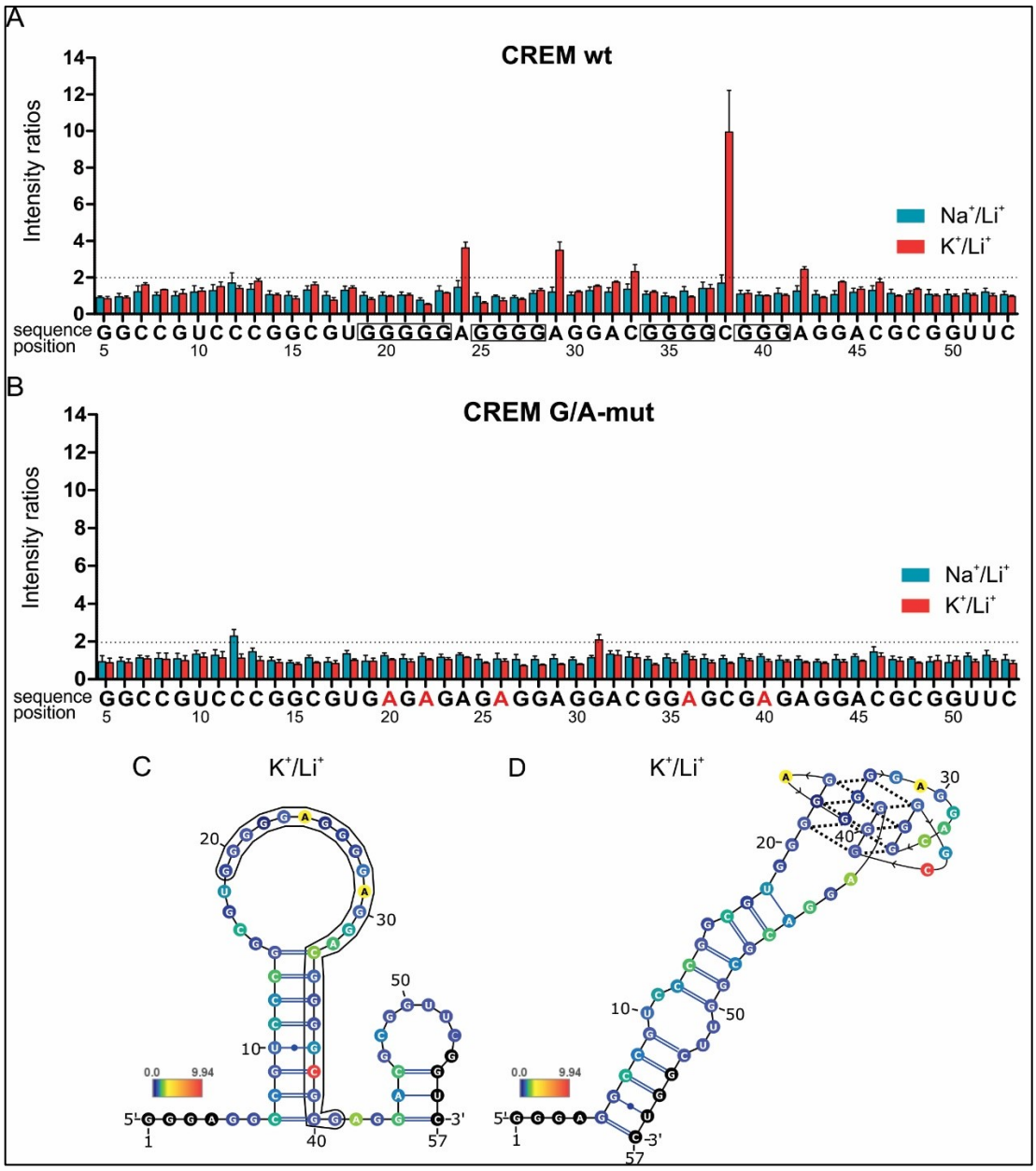



**Figure 16** – Semi-quantitative analysis of the in-line probing experiments and interpretation of the secondary structures.

(A–B) Ratios of the bands' intensity of the CREM wt (A) and G/A-mut (B) *in vitro* PG4 versions for each nucleotide. The $Na^+/Li^+$ ratios are shown in blue and $K^+/Li^+$ ones are in red. The dotted lines represent the 2-fold threshold that denotes a significant gain in flexibility. Both the sequence and the positions of the nucleotides are indicated on the *x*-axis. The boxed guanines represent the G-tracts involved in the G-quadruplex formation. The adenines shown in red are those replacing the guanines in the CREM G/A-mutant. Each bar represents the average of three independent experiments, and the error bars represent the standard deviations. (C–D) The $K^+/Li^+$ ratios of bands' intensity transposed as a color map on the best predicted secondary structure of the CREM wt PG4 either *in vitro* (C), or on the predicted structure containing the additional 5' and 3' regions flanking the PG4 with the putative CREM G-quadruplex folded at the top of the stem (D). The flexibilities of the nucleotides

are proportional to their ratio of bands' intensity. Low ratios are shown in blue and high ones in red. The boxed nucleotides represent the predicted PG4 region. The nucleotides shown in black are those for which no significant ratio could be calculated because their representative bands either migrated off of the gel, or were not sufficiently resolved in the electrophoresis.

These various representations (see **Figures 15 and 16**) facilitate the identification of the secondary structure that most likely fits the in-line probing data obtained under conditions either unfavorable or favorable for the formation of the G-quadruplex.

## 3. RESULTS AND DISCUSSION

### 3.1. Molecular design

In the last few years, RNA G-quadruplexes found in the 5'-UTRs of mRNAs acting as translational repressors have attracted a lot of attention (for a review see (Bugaut et Balasubramanian, 2012)). In this vein, a PG4 sequence found in the 5'-UTR of the human CREM mRNA was chosen with which to illustrate, step-by-step, an in-line probing protocol that analyzes the ability of this candidate to fold *in vitro* into a G-quadruplex structure. Multiple tools permit the prediction of G-quadruplex formation (Cer *et al.*, 2012; Kikin *et al.*, 2006; Scaria *et al.*, 2006; Wong *et al.*, 2010), and various databases of the PG4 sequences found in pre-mRNAs, mature mRNAs and both 5'- and 3'-UTRs are publically available (Beaudoin et Perreault, 2010 ; Huppert *et al.*, 2008 ; Kikin *et al.*, 2008). The content of most of these databases was generated using the algorithm for predicting PG4 motifs mentioned in Section 2.1 (Materials and Methods). The CREM PG4 was chosen from a database built in our laboratory, and is located in the 5'-UTR of the human CREM transcript variant 19 mRNA (NM_183013) (Beaudoin et Perreault, 2010). This 5'-UTR is 407 bp long and the PG4 sequence starts at position 110. The CREM PG4 is predicted to be composed of 23 nt (**Figure 13A**), to possess several relatively short loops and to not contain any important cytosine stretches in its flanking sequences. These characteristics strongly increase its probability of folding into a G-quadruplex structure both *in vitro* and *in cellulo*. Indeed, these two criteria have been demonstrated to greatly influence not only the ability of an RNA sequence to fold into a G-quadruplex structure, but also its stability (Beaudoin et Perreault, 2010 ; Zhang *et al.*, 2011a). Clearly, the CREM PG4 represented an ideal candidate for this study. A previous study showed that it was more representative of the actual cellular context to probe, *in vitro*, a slightly extended version of the PG4 sequence in question in order to

obtain more accurate data (Beaudoin et Perreault, 2010). Consequently, the sequence probed included an additional 18 nt at the 5'-end and 16 nt at the 3'-end of the PG4 sequence and is referred to as *in vitro* PG4 (**Figure 13A**). The added nucleotides were identical to those found in the natural 5'-UTR. Finally, a mutant version in which several guanines were substituted for by adenines was also synthesized. These substitutions have the effect of disrupting the G-tracts and, consequently, abolishing the ability of the RNA to fold into a G-quadruplex structure (see the lower case "g" in **Figure 13A** corresponding to the G/A-mutations). We suggest to use this handbook for the mutation of G-tracts in order to make sure to disrupt them adequately: GGG/GaG; GGGG/GaGG or GGGG/GGaG; GGGGG/GaGaG; GGGGGG/GGaGaG or GGGGGG/GaGGaG; and so on.

### 3.2. Secondary structure predictions

Previous characterizations of many PG4 sequences revealed that some do not in fact adopt a G-quadruplex structure because, instead, they fold into stable secondary structures that are formed by Watson–Crick base pairs (Beaudoin et Perreault, 2010, and unpublised data). Because the latter structures are rapidly formed, this significantly impairs the folding into a G-quadruplex structure, a process that requires more time both *in vitro* and *in vivo*. As a result of these observations, the protocol was adapted so as to consider the predicted secondary structures based on both Watson–Crick base pairs and on the one including the G-quadruplex for each of the *in vitro* PG4 candidates studied. The secondary structures of the designed wt CREM *in vitro* PG4 version was predicted using the RNAstructure software (Reuter et Mathews, 2010). Two potential structures were obtained (**Figure 13B and C**). Briefly, the first one includes two hairpins, of 6 and 7 base pairs that are linked by two single-stranded nucleotides and harbours medium sized loops (**Figure 13B**). The second is also composed of two hairpins, which are different from those of the previous structure, and are linked by three single-stranded nucleotides. Here, the first hairpin is composed of an 8 base pair stem that is capped by a large 18 nt loop, while the second is a small one composed of a 3 base pair stem and harbouring a 6 nt loop (**Figure 13C**). One possible way to differentiate both of these predicted structures based on in-line probing should be the position of the single-stranded regions in both structures as they are distinct. Importantly, both of the predicted secondary structures showed a limited stability which was estimated to be between -18.8 and -17 kcal/mol (**Figure 13B and C**, respectively).

When looking at the probability of a given sequence to form a G-quadruplex structure, an intrinsic parameter should be the number of nucleotides of the potential PG4 sequence that might be involved in Watson–Crick base pairs according to the RNA structure prediction. In the case of the CREM PG4 sequence, 10 of the 23 nt appeared to be in single-stranded regions in the first predicted structure as compared to 15 nt in the second structure (**Figure 13B and C**). Considering both the lack of highly stable predicted secondary structures, and the relative abundance of single-stranded nucleotides, the CREM PG4 appeared to be a suitable candidate to fold into a G-quadruplex.

RNAstructure folding software cannot predict the presence of a G-quadruplex motif. Nonetheless, the folding of the nucleotides on either side of the PG4 was predicted by preventing the PG4 region from being involved in the folding. In order to do so, the predicted unimolecular parallel G-quadruplex was considered as being already folded and was removed from the equation (**Figure 13D**). The sequences surrounding the PG4 (i.e. the 5'- and 3'-extensions) were then folded together, if possible as a helical region (**Figure 13D**). For the CREM wt sequence, this permitted the formation of an additional stem of 5 base pairs.

### 3.3. RNA synthesis and in-line probing

Subsequent to the designing of the sequence and the analysis of their predicted computer-based structures, RNA transcripts have to be synthesized. Double-stranded DNA templates for both the wt and G/A-mut versions of the CREM candidate were synthesized by the filling of two partially complementary oligonucleotides (**Figure 12**). Upon performing the experiment it was noticed that DMSO was generally essential for this step. It creates slightly denaturing conditions that impair stable secondary structure formation, and thus permit the polymerase to read through the entire sequence. DMSO is known to increase the PCR amplification efficiency of GC-rich sequences (Varadaraj et Skinner, 1994). Once the DNA templates were ready, they were *in vitro* transcribed using purified T7 RNA polymerase (see Section 2.3). The resulting reaction mixtures were treated with DNase to remove the DNA template. Phenol chloroform extraction was then performed in order to remove the proteins, and, lastly the RNA transcripts were fractionated by denaturing (8 M urea) 10% polyacrylamide gel electrophoresis. The RNAs in the gel bands of the appropriate sizes were recovered, dephosphorylated and 5'-end labelled with $^{32}$P using standard procedures.

Prior to the in-line probing experiment, trace amounts of all RNA samples were denatured at 70°C for 5 min, followed by slow-cooling to room temperature in the presence of 100 mM of monovalent cation (i.e. either $Li^+$, $Na^+$ or $K^+$). In principle, this step should favor the prefolding of G-quadruplexes or other RNA structures. After the addition of the in-line probing buffer, all RNA samples were subjected to in-line probing reactions at room temperature for 40 h. The length of the incubation should be sufficient for the G-quadruplex structure to be formed and to reach equilibrium. Hydrolysis of the phosphodiester bonds was observed to occur in the most flexible regions.

It is important to note that only a trace amount of RNA (50 000 cpm, < 1 nM) is characteristically used in the in-line experiment. Therefore, most likely only intramolecular G-quadruplex formation is possible. This is an important difference as compared to other biophysical methods that are commonly used to study G-quadruplex formation, as methods such as circular dichroism and thermal denaturation require RNA concentrations in the low micromolar range which permit the formation of intermolecular G-quadruplexes. In our opinion, limiting the analysis to solely unimolecular topologies by using trace amounts of RNA is more biologically relevant, and is therefore essential in order to be able to properly evaluate both the potential of G-quadruplex formation and the role of these structures *in cellulo*. Although, even if it has never been observed in our hand, the relatively high concentration of magnesium ions (20 mM) could potentially affect the RNA structure equilibrium (e.g. between G-quadruplexes and alternative secondary structures) represents a limit of the technique. To get over this possible limitation, we suggest to confirm in-line probing results with a short experiment using one of the complementary biophysical methods mentioned above in condition corresponding to physiological concentration of magnesium ions (~1 mM).

After the incubation period, equivalent amounts of radioactivity (cpm) from each reaction were analyzed on a denaturing polyacrylamide gel. The bands were visualized by exposure of the dried gel to a phosphorscreen. A typical autoradiogram for both the wt and G/A-mut versions of the CREM candidate probed in the presence of either $Li^+$, $Na^+$ or $K^+$ is shown in **Figure 14**. A change in the banding patterns was observed solely for the wt sequence. More precisely, specific nucleotides appeared to become more susceptible to hydrolysis in the presence of $K^+$. It is noteworthy that the $Li^+$ cation is an excellent negative

control in the study the formation of G-quadruplexes as it maintains the same ionic strength in solution, but is unable to stabilize the stacking of the G-quartets due, primarily, to its smaller size. In other words, it favors the formation of the Watson–Crick base pair based secondary structure. Conversely, the presence of $K^+$ may stabilize both the G-quartet motifs and their stacking, which, therefore, favours the formation of a G-quadruplex structure. The bands showing an increased intensity in the presence of $K^+$ correspond to those nucleotides located within the predicted loops that are intercalated between the guanosine tracts, as well as those located immediately 3' of the PG4 sequence (**Figure 14**; nucleotides $A_{24}$, $A_{29}$, $C_{33}$, $C_{38}$ and $A_{42}$). All of these regions were predicted to be single-stranded and therefore are probably more flexible upon formation of the G-quadruplex, thereby supporting the folding into this structure. Contrastingly, the susceptibility to hydrolysis of the corresponding nucleotides in the G/A-mut version remained unchanged when probed in the presence of $K^+$ instead of $Li^+$. Finally, the same probing pattern was observed in the presence of $Li^+$ and $Na^+$, suggesting that no G-quadruplex structure was formed by this sequence in presence of $Na^+$.

### 3.4. Semi-quantitative analysis of the in-line probing

In order to achieve a more robust analysis, and to provide a quantitative aspect to the probing, triplicate experiments of in-line probing reactions were performed for each RNA sequence. The resulting band intensities were then quantified for each band using the SAFA Software (Laederach *et al.*, 2008). The $K^+/Li^+$ intensity ratio was calculated for each position for both the wt and the G/A-mutant versions. The average and standard deviation (SD) of these ratios for each position and sequence were used to build bar graphs. Examples for the CREM candidate are illustrated in **Figure 16A and B**. In order to determine if a specific nucleotide was truly more accessible in the presence of $K^+$ as compared $Li^+$, the $K^+/Li^+$ ratio was compared to that of the G/A-mutant. The reproducibility of the results is illustrated by the analysis of the G/A-mut sequence, which should exhibit no structural difference between these two ionic conditions, and thus permits the establishment of a threshold. In fact, no significant variation was observed in the bands' intensities between the three ionic conditions for the G/A-mutant version (**Figure 16B**). The ratios over this threshold value should represent those nucleotides that are significantly more flexible. The study of more than twenty G-quadruplexes (Beaudoin et Perreault, 2010, and unpublished data) indicated that a threshold of 2-fold was an accurate indication of a nucleotide that shows a significantly

higher flexibility. For the CREM wt sequence, five nucleotides showed $K^+/Li^+$ ratio over 2 ($A_{24}$, 3.61; $A_{29}$, 3.48; $C_{33}$, 2.32; $C_{38}$, 9.94; and $A_{42}$, 2.43) (**Figure 16A**). Four of these are located in the predicted loops of the folded G-quadruplex. More specifically, these nucleotides are situated immediately either 5' or 3' of the G-tracts. The last of the fare is located at the 3' end of the last G-tract (i.e. in position 42). According to our other probings of RNA G-quadruplexes, this is typical. Depending on the particular G-quadruplex studied, the sequences on both the 5' and 3' sides of the PG4 region can also be affected by the G-quadruplex's formation (Beaudoin et Perreault, 2010, and unpublished data). Moreover, it was observed that pyrimidine residues (i.e. C and U) are more susceptible to exhibiting significant hydrolysis in the G-quadruplex structure, in good agreement with a previous demonstration that pyrimidines are more prone to non-enzymatic spontaneous hydrolysis than are purines (Li et Breaker, 1999). This might explain why some nucleotides in the loop, -GGA- in loop 2 of CREM for instance, showed superior cleavage levels ($K^+/Li^+$ ratio of 1.21, 1.52 and 1.74 respectively), but remain under the fixed 2-fold threshold (**Figure 16A**). In summary, a clear modification in RNA structure driven by the presence of $K^+$ was observed. Moreover, the new structural features seemed to support the folding into a G-quadruplex structure. Finally, this procedure brings a semi-quantitative aspect to the analysis; however, it should always be considered with precaution and the appropriate controls must always be performed.

Several additional controls were required in order to validate the method, more specifically to verify that the quantification and the ratio calculations were accurate. Firstly, the amount of radioactivity of all of the samples was determined and equivalent amounts of cpm were loaded onto the gel for each of the samples. After migration and visualization by phosphorimaging, the total amounts of cpm in each of the lanes containing the in-line probing samples were quantified (using the ImageQuant software version 7.0; GE Healthcare Life Sciences) and compared (**Supplementary Figure 17, S1**). For each gel, the average radioactivity, in terms of cpm, for all of the lanes was calculated. If the standard deviation was too high (±15%), the results of a specific lane, or the complete gel, were rejected. This event in fact occurred very unfrequently. Secondly, equal amounts of cpm of a specific CREM wt PG4 sample were loaded into two distinct wells in order to assess any possible bias arising from either the loading step or the position of the samples on the gel. No

significant variation was observed, in terms of cpm, between the intensities of the bands in the two lanes, nor in the banding patterns (**Supplementary Figure 18, S2**). Thirdly, since the $K^+/Li^+$ ratios used for building the bar graphs represent the averages of three distinct experiments, standard deviations (SD) were determined and are illustrated using error bars (see **Figure 16A and B**). Clearly, the standard deviations were relatively small. Finally, this method was applied to several other candidates in order to ensure that it worked for candidates other than CREM. Specifically, more than twenty transcripts including potential G4 structures have been probed to date and in all cases conclusive data were obtained (Beaudoin et Perreault, 2010, and unpublished data).

### 3.5. Comparing structure predictions and *in vitro* probing results

With the results of the in-line probing experiments in hand, it is of interest to take a closer look at the secondary structure adopted by the transcripts, starting with the one found in presence of $Li^+$ (i.e. under conditions unfavorable to G-quadruplex formation). In order to do so, the raw intensity values of the $Li^+$ conditions were specifically normalized with the help of a methodology used for the analysis of SHAPE results (see Section 2.6). The results of this normalization correspond to ratios that represent the levels of cleavage for each nucleotide under the same conditions (in the presence of $Li^+$ here). An initial color code can be created with these ratios, and the values can be superposed on the two initial predicted secondary structures (**Figure 13B and C**). Blue indicates constrained (base-paired) nucleotides, and colors from green to yellow to red indicate regions of increasing flexibility and accessibility, that is to say residues that are most likely single-stranded (or are less stable). Clearly, the best fit was with the second predicted structure that is composed of two stem loops with loops of 18 and 8 nt (compared **Figure 15B to A**). Specifically, the in-line probing results showed that the most accessible nucleotides were found to be located in the hairpin loops, and the long stem was confirmed to contain constrained nucleotides (**Figure 15B**). Thus, the experimental data support the second predicted structure for the CREM wt sequence under conditions unfavorable for G-quadruplex formation (i.e. in the presence of $Li^+$). A second color code can be produced using the averaged $K^+/Li^+$ ratios presented above (Section 3.4). With this new code, blue represents a ratio near 1, and colors from yellow to red show increasing ratios up to 9.94, the maximum ratio observed for the CREM PG4. The results of this second color code were transposed onto the secondary

structure suggested to be adopted in the presence of $Li^+$, as well as on that obtained with the predicted unimolecular parallel G-quadruplex structure (**Figure 16C and D**). With these representations, it appeared obvious that the differences in accessibility in the presence of the $Li^+$ versus in the presence of $K^+$ preferentially occurred for the nucleotides located in the loops and in those located 3' of the PG4 region. Clearly, the $K^+/Li^+$ values have a better fit on the predicted structure that includes the G-quadruplex (**Figure 16D**), as several discrepancies are observed when the structure including solely Watson–Crick base pairs is considered (**Figure 16C**). In summary, the in-line probing of the CREM wt transcript unambiguously demonstrated the transition from a secondary structure composed of two stem loops to a unimolecular G-quadruplex structure is due to the presence of KCl.

## 4. CONCLUDING REMARKS

The in-line probing method appears to be a simple, robust, reproducible and informative one with which to study RNA G-quadruplex formation. More importantly, compared to circular dichroism, thermal denaturation and NMR techniques, a much lower concentration of RNA is required for in-line probing (i.e. <1 nM). Specifically, only trace amounts of RNA are necessary, which permits avoiding the potential formation of intermolecular G-quadruplexes. Another advantage is that it is relatively quick to perform, as only a few days are required for both the probing and the analysis of both the wt and the mutated versions.

As presented, in-line probing permits the confirmation of whether or not a given PG4 sequence folds into a G-quadruplex structure. The corresponding G/A-mutant version does not permit this folding and is in fact important when further *in cellulo* investigations of the G-quadruplex need to be performed. Moreover, in-line probing offers the advantage of providing information on the structural modifications of the whole molecule following the G-quadruplex's formation. It is also possible to gain structural information for the nucleotides located on both sides of the G-quadruplex motif. So far, we have successfully used this technique to probe G-quadruplex sequences found in RNA molecules over 120 nt long (unpublished data). The possibility to probe relatively long molecule may be instructive in several situations, for example if the formation of a G-quadruplex is used to expose an adjacent regulatory region that was previously trapped in a hairpin, or the opposite situation, in which it is used to hide a region that was previously accessible.

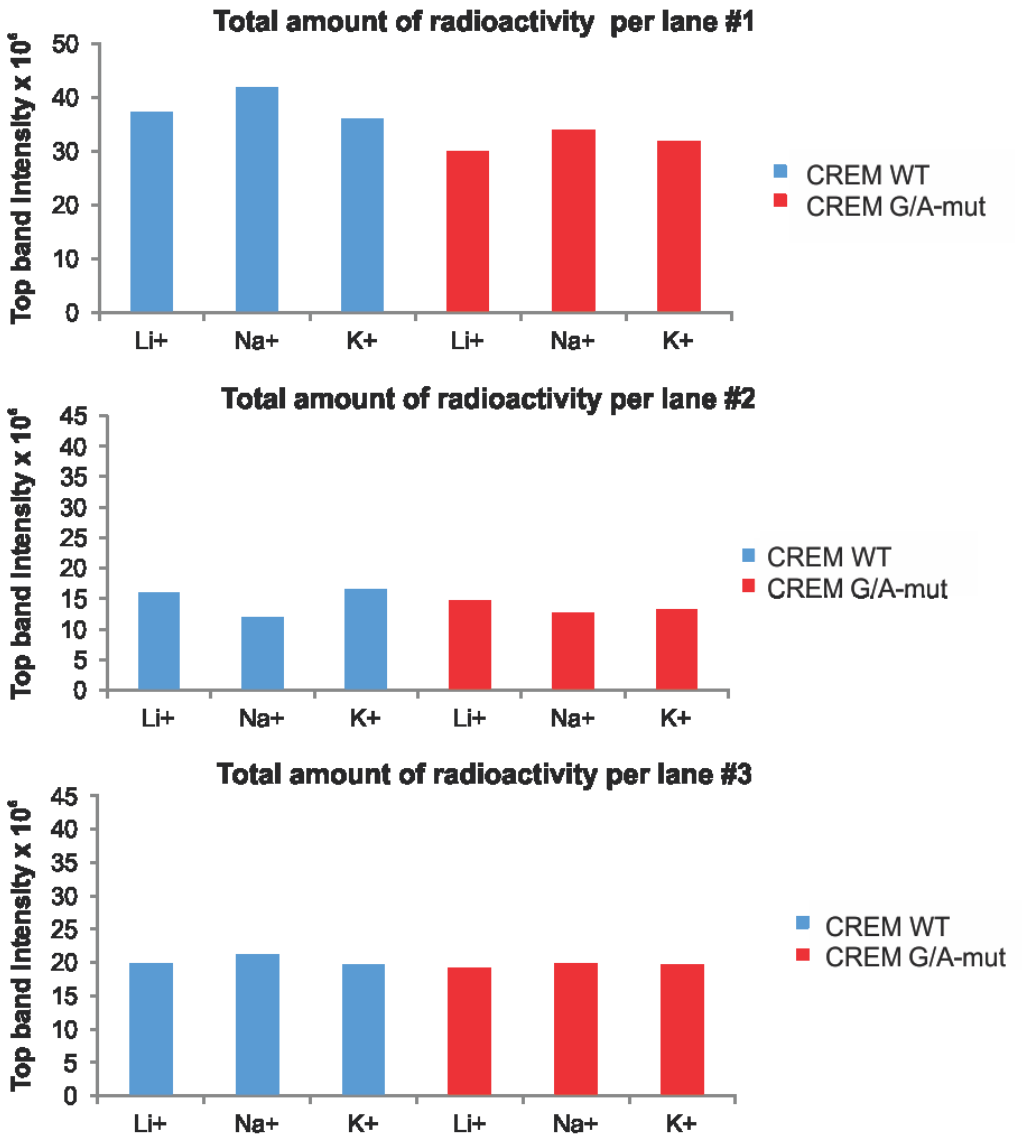
In brief, the detailed methodology described here combines the use of bioinformatic algorithms to identify potential G-quadruplex sequences, a program for secondary structure prediction, in-line probing and its semi-quantification analysis and the representation of the resulting structure. Together, this represents a complete and accurate method with which to study RNA G-quadruplex formation. The results obtained are easy to interpret and provide a concrete and understandable visualization of the various structures adopted in different conditions.

## ACKNOWLEDGMENTS

This work was supported by a grant from the Canadian Institute of Health Research (CIHR, grant number MOP-44022) to J.P. Perreault. The RNA group is supported by a grant from the Université de Sherbrooke. J.D.B. was the recipient of the CIHR Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Award. R.J. was the recipient of the CIHR Frederick Banting and Charles Best Canada Graduate Scholarship Master's Award. J.P.P. held the Canada Research Chair in Genomics and Catalytic RNA and currently holds the Chaire de Recherche de l'Université de Sherbrooke en Structure et Génomique de l'ARN. He is also a member of the Centre de Recherche Clinique Étienne-Le Bel.

**SUPPLEMENTARY DATA**

A

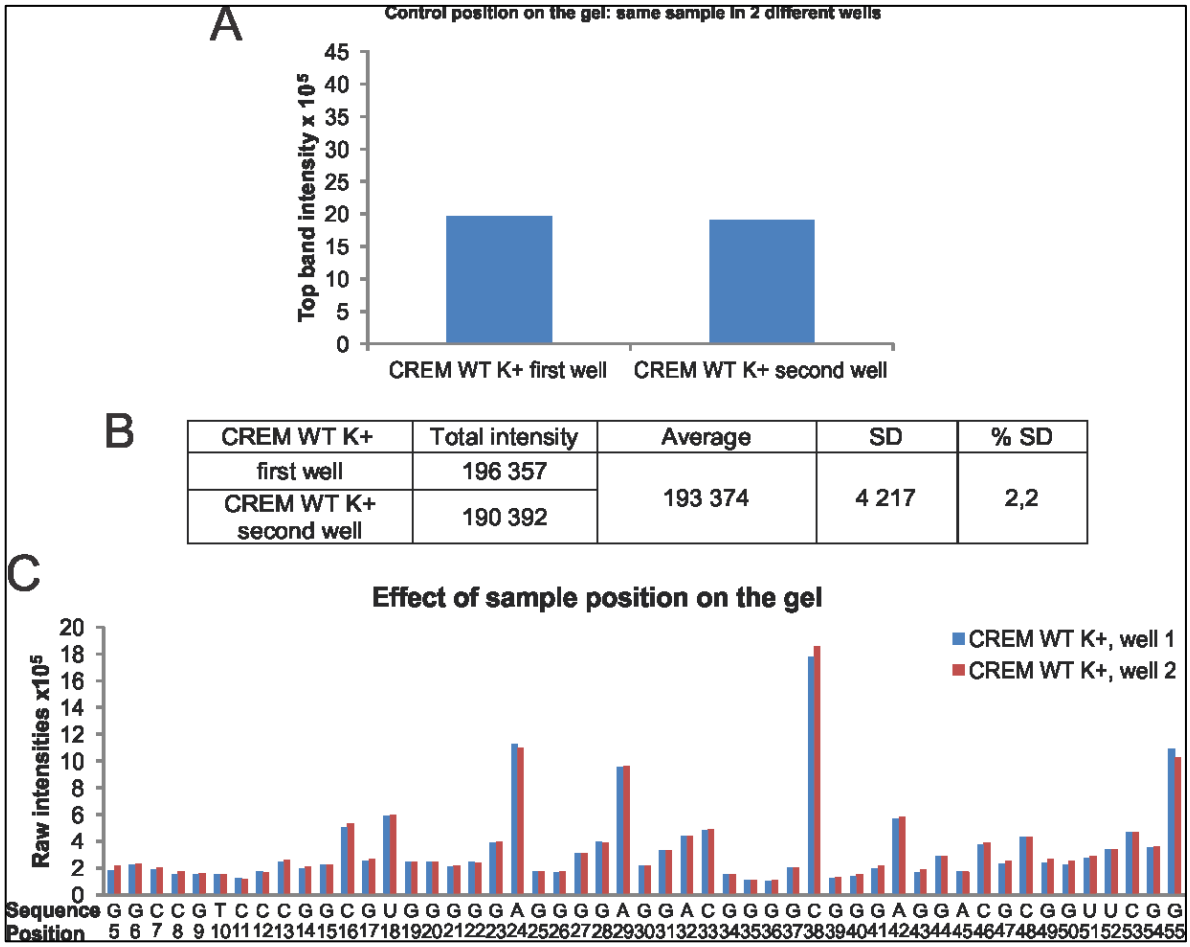



B

| | average intensity | SD | % SD |
|---|---|---|---|
| Gel #1 | 350 791 | 41 859 | 11.9 |
| Gel #2 | 141 314 | 18 329 | 13.0 |
| Gel #3 | 198 030 | 7 144 | 3.6 |

**Figure 17** – S1 Total amount of radioactivity in each lane of the gels.

(A) Bar graphs showing the top band intensities (quantified with ImageQuant) of every lane for the triplicates gels. This top band band is representative of the total amount of radioactivity present in the lane. The blue bars are the CREM wt PG4, and the red ones the CREM G/A-mut PG4 in all the three conditions tested. (B) Table showing the average total radioactivity for all of the samples of the same gel with their standard deviations (SD).




| CREM WT K+ | Total intensity | Average | SD | % SD |
|---|---|---|---|---|
| first well | 196 357 | 193 374 | 4 217 | 2,2 |
| CREM WT K+ second well | 190 392 | | | |

**Figure 18** – S2 Effect of the sample position on the gel on the intensity measurements.

(A) Bar graph showing the top band intensity (quantified with ImageQuant) of the same sample (CREM wt $K^+$, replicate #3) loaded into two different wells on the same gel. (B) Table of the total intensity, average and standard deviations of the CREM wt $K^+$ sample loaded into two different wells of the same gel. (C) Raw intensities for every nucleotide of the same sample (CREM wt $K^+$, replicate #3) loaded into two different wells of the same gel (the first well is in blue, the second in red). The same intensities are observed for the corresponding nucleotides, and, overall, an identical pattern of band intensity is observed. The sequence and positions of the nucleotides are indicated on the X-axis.

# ARTICLE 2 – NEW SCORING SYSTEM TO IDENTIFY RNA G-QUADRUPLEX FOLDING

**Auteurs de l'article** : Beaudoin, Jean-Denis*, Jodoin, Rachel* et Perreault, Jean-Pierre
* Co-premiers auteurs

**Statut de l'article** : Publié dans Nucleic Acids Research (2014) vol. 42, no. 2, p.1209-1223

**Avant-propos** : Rachel Jodoin a effectué les expériences de cartographie *in vitro* des candidats avec long contexte ainsi que des 14 séquences pour tester le score final. En collaboration égale, Rachel Jodoin et Jean-Denis Beaudoin ont effectué les analyses comparatives des scores et les essais *in cellulo*. Jean-Denis Beaudoin est l'instigateur d'un concept de score pour la prédiction de G4. Il a effectué les expériences initiales de cartographie, de dichroïsme circulaire et de dénaturation thermique sur le candidat TTYH1. Le manuscrit a été rédigé par Rachel Jodoin, Jean-Denis Beaudoin et Jean-Pierre Perreault.

**Résumé**

Les G-quadruplexes (G4) sont des structures non canoniques impliquées dans plusieurs processus cellulaires d'importance. À ce jour, la prédiction des structures G4 potentielles (PG4) est basée presque exclusivement sur la séquence d'intérêt respectant l'algorithme $G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$ (où $x \geq 3$ et N=A, U, G ou C). Cependant, plusieurs séquences respectant cet algorithme ne forment pas de G4 et sont considérées comme des prédictions faussement positives. Dans cette étude, nous démontrons que le candidat PG4 d'ARN situé dans la région 3'UTR du gène TTYH1 est l'un de ces faux positifs. Spécifiquement, le repliement G4 a été observé comme étant inhibé par la présence d'une série de cytosines consécutives situées dans le contexte génomique du candidat, résultant en l'adoption d'une structure Watson-Crick canonique. De toute évidence, les séquences avoisinantes des PG4 peuvent influencer leur repliement. La structure secondaire a été évaluée par cartographie *in-line* pour 12 motifs PG4, entourés en amont et en aval par 15- ou 50-nt provenant de leur contexte génomique. Ces données ont permis le développement d'un système de score pour

la prédiction des PG4 qui considère les séquences avoisinantes. L'exactitude de ce système de score a été testée par la cartographie de 14 nouveaux candidats PG4 retrouvés dans des séquences 5'UTR humaines. Ce nouveau système de score, en combinaison avec l'algorithme de recherche standard, peut être utilisé afin de mieux prédire les repliements des G4 d'ARN.

**Abstract**

G-quadruplexes (G4s) are non-canonical structures involved in many important cellular processes. To date, the prediction of potential G-quadruplex structures (PG4s) has been based almost exclusively on the sequence of interest agreeing with the algorithm $G_x$-N-$_{1-7}$-$G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$ (where x≥3 and N=A, U, G or C). However, many sequences agreeing with this algorithm do not form G4s and are considered false positive predictions. Here we show that the RNA PG4 candidate present in the 3'-untranslated region (UTR) of the TTYH1 gene to be one such a false positive. Specifically, G4 folding was observed to be inhibited by the presence of multiple-cytosine tracts, located in the candidate's genomic context that adopted a Watson-Crick base-paired structure. Clearly, the neighbouring sequence of a PG4 may influence its folding. The secondary structure of 12 PG4 motifs along with either 15 or 50 nucleotides of their upstream and downstream genomic contexts were evaluated by in-line probing. Data permitted the development of a scoring system for the prediction of PG4s taking into account the effect of the neighbouring sequences. The accuracy of this scoring system was assessed by probing 14 other novel PG4 candidates retrieved in human 5'-UTRs. This new scoring system can be used, in combination with the standard algorithm, to better predict the folding of RNA G4s.

## INTRODUCTION

Guanine-rich nucleic acid sequences can fold into a non-canonical tetrahelical structure termed a G-quadruplex (G4). The primary building block of this structure, the G-tetrad, is composed of four co-planar guanines that interact with each other via Hoogsteen base pairs and are stabilized by a metal cation, usually potassium. The stacking of these G-tetrads forms a G4, which is a stable structure. Several bioinformatics studies reported enrichment in the number of potential G-quadruplex (PG4) sequences found in various DNA and RNA regulatory elements, respectively, located within the genome and the transcriptome (Huppert *et al.*, 2008; Huppert et Balasubramanian, 2005; Todd *et al.*, 2005). Promoters, telomeres and both the 5'- and 3'-untranslated regions of mRNA (UTRs) are some examples of these elements. Recently, an elegant approach based on an engineered, structure-specific antibody led to the direct quantitative visualization of DNA G4s inside living cells (Biffi *et al.*, 2013). This study demonstrated that G4 formation in the nucleus of cells was modulated during cell-cycle progression, and that endogenous DNA G4s can be stabilized by a small-molecule ligand. Even though a quantitative, direct visualization of RNA G4 structures inside living cells is still lacking, over the last few years several roles have been attributed to RNA G4s [for a review see (Millevoi *et al.*, 2012)]. These include: pre-mRNA splicing and polyadenylation, mRNA translation and targeting, transcriptional termination and telomere homeostasis (Millevoi *et al.*, 2012). Clearly, RNA G4s appear to be one of the key motifs of the transcriptome. Thus, learning to accurately predict and locate RNA G4s is crucial to unlocking the study of their biological functions and impacts.

So far, most studies of biological G4 structures have combined bioinformatics predictions supported by physical evidence of G4 folding *in vitro*, as well as assessment of potential biological roles in cell culture assays , for examples see (Beaudoin et Perreault, 2010, 2013 ; Halder *et al.*, 2012) . A key step in this investigative process is of course the initial prediction of G4 folding. This is almost exclusively based on the computerized identification of potential G4 (PG4) sequences using a specific search algorithm (or close derivatives thereof) for the $G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$, sequence where $x$ is $\geq 3$ and $N$ corresponds to any of the four nucleotides (A, G, C and T or U)(Huppert et Balasubramanian, 2005 ; Todd *et al.*, 2005). These algorithm criteria were developed using results from various *in vitro* experiments but primarily from DNA G4 folding studies. Several discrepancies

concerning PG4s identified via this algorithm have been reported in recent years. Certain sequences not fulfilling all of the algorithm's criteria were indeed shown to fold into G4s that is to be false negatives. The DNA G4 reported for the CEB25 minisatellite is a good example of one such false negative (Amrane *et al.*, 2012). Because of its central 9-nt loop, the algorithm did not predict it would form a G4. It has also been shown that DNA PG4s including single-nucleotide loops 1 and 3 support the presence of a large loop 2 of up to 21 nt in length (Guédin *et al.*, 2010). Similarly, RNA G4s including loops up to 15 nt long have also been reported to fold into stable G4s, both *in vitro* and *in cellulo* (Pandey *et al.*, 2013) (Rouleau S., Beaudoin JD. and Perreault JP. unpublished data). Recently, Mukundan and Phan reported the *in vitro* formation of artificial DNA G4s with multiple bulges involving discontinuous guanine tracts (G-tracts) that is differing from the standard algorithm (Mukundan et Phan, 2013). Another guanine-rich RNA sequence ignored by the algorithm was reported to form an atypical G4 bearing discontinuous G-tracts. The high-resolution structure determined for the *sc1* RNA bound to a peptide from the human fragile X mental retardation protein eloquently illustrates both the heterogeneity and complexity of the web of RNA strand interactions involved in G4 folding (Phan *et al.*, 2011). There are also many reports of false positives (i.e. PG4s identified via the algorithm that do not to fold into G4s) both *in vitro* and *in cellulo*. One study focusing on human 5'-UTR mRNA G4s reported that several selected PG4s fulfilling all of the algorithm's requirements were in fact unable to fold into G4s (Beaudoin et Perreault, 2010) owing to cytosines tracts (C-tracts) located in their flanking sequences – that is within 10–15 nt either in 5' and 3' of the PG4. It turns out that these C-tracts interacted with the G-tracts of the PG4 sequence, producing alternative secondary structures based on Watson-Crick base pairs. This impaired G4 folding by sequestering key guanines. For each of these non-folding PG4s, substitution mutants bearing adenosines instead of cytosines, to destabilize the inhibitory secondary structure, were shown to successfully fold into G4s (Beaudoin et Perreault, 2010).

Here, we investigated the folding of a PG4 located in the 3'-UTR of the TTYH1 gene both *in vitro* and *in cellulo.* Data indicate that the presence of multiple cytosines within the PG4's genomic context inhibits G4 folding. To broaden our investigation of the influence of the PG4's neighbouring sequences and their impact on G4 folding, we screened multiple biological RNA PG4s. Results permitted the development of a predictive score for G4

folding. This novel scoring system can be used to cure PG4 databases of false-positive candidates. The risks and benefits of our scoring system for the identification of PG4s within genomes and transcriptomes are also discussed.

## MATERIALS AND METHODS

### Bioinformatics

Human 5'- and 3'-UTR databases were derived from sequences obtained from UTRdb (UTRef release 9) (Mignone *et al.*, 2005). PG4 sequences were ascertained using the RNAMotif program (Macke *et al.*, 2001) and the following algorithm search sequence: $G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$, where $x$ is ≥ 3 and $N$ corresponds to any of the 4 nt (A, G, C or U). Only PG4s distanced by a minimum of 10 nt were retained. Data output from the RNAMotif program was exported into Excel file format using various Perl scripts to generate the data sets. Data sets 1 and 2 from 5' and 3'-UTRs respectively are available in the Supplementary Material. Minimum free energy values (Mfe, kcal/mol) were predicted by the RNAfold software from the Vienna RNA Package (Hofacker *et al.*, 1994).

The cG score is calculated on a string 's' of length 'n' as follows:

'Gs(i). represents the set of all substring of consecutive 'Gs. found in s, and '|Gs(i)|' is the cardinality of this set. Note that all substrings in this set are identical but correspond to different regions of 'S' (e.g. the string 'GGGCGGG' has 2 'GGG' substrings and thus |Gs (3)| will be 2)

The cG score of string s is then defined as

$$cG(s) = \sum_{i=1}^{n} (|Gs(i)| * 10 * i)$$

The cC score is calculated in a similar manner:

$$cC(s) = \sum_{i=1}^{n} (|Cs(i)| * 10 * i)$$

In other words, for a given PG4, a value of 10 is attributed for each G or C, and then, a value of 20 for each doublet (GG or CC), a value of 30 for each triplet (GGG or CCC), and so on. The cG or cC score is the sum of all Gs' and Cs' attributed values respectively. For example, three consecutive Gs will generate a total cG score of 100 because it is counted as three single Gs, two different doublets and one triplet [$cG\ score = 3\ (G) \times 10 + 2(GG) \times$

$20 + 1(GGG) \times 30 = 100$ ], whereas two consecutive Gs have a total cG score of 40[($cG\ score = 2\ (G) \times 10 + 1(GG) \times 20 = 40$].

Finally, the cG/cC score was calculated as the ratio of both scores:

$$cG/cC\ score = \frac{cG\ score}{cC\ score}$$

Receiver-operator characteristic (ROC) curves analyses were performed using the GraphPad Prism version 5.02 for Windows (GraphPad Software, San Diego California USA). Briefly, total loop length, Mfe, cG/cC score and QGRS G-score values of each long-context PG4 (12 candidates of the first set) were divided into G4-folding or non-folding categories, based on in-line probing results. Specificity was measured as the fraction of non-folding candidates with prediction parameters inferior to the threshold value. Sensitivity was evaluated as the fraction of G4-folding candidates with a predictive parameter over the threshold value. The threshold was the value midway between a pair of PG4 values. Paired specificities and sensitivities were evaluated for all such threshold values and plotted on a graph where the area under the curve (AUC) is the ability to discriminate between G4-folding and non-folding. An AUC of 0.5 is a random (i.e. non-discriminating value), whereas an AUC of 1 demonstrates perfect discrimination.

**RNA synthesis**

The detailed protocol for the analysis of PG4s by in-line probing has been described previously (Beaudoin *et al.*, 2013). First, double-stranded DNA sequences corresponding to each PG4, and containing the T7 RNA promoter sequence, were prepared. All of the oligodeoxyribonucleotide sequences used are presented under **Supplementary Table S1 in Annexe 2**. Two overlapping oligonucleotides (2 µM each, Invitrogen) were annealed and then purified *Pfu* DNA polymerase was used to fill in the gaps in the presence of 5-10% DMSO (Fisher scientific), 2 mM $MgSO_4$, 0.2 mM of each dNTP, 20 mM Tris-HCl, pH 8.8, 10 mM KCl, 10 mM $(NH_4)SO_4$ and 0.1% Triton X-100. Full length double-stranded DNAs were then ethanol-precipitated and the resulting pellet dissolved in ultrapure water. RNA transcripts were prepared by *in vitro* run-off transcription using purified T7 RNA polymerase (10 µg) in the presence of 20 U of RNase OUT (Invitrogen), 0.01 U of pyrophosphatase (Roche Diagnostics) and 5 mM NTPs in a buffer containing 80 mM HEPES-KOH, pH 7.5, 24 mM $MgCl_2$, 2 mM spermidine and 40 mM DTT, all in a final

volume of 100 µL. The reactions were incubated at 37°C for 2 h. Fifteen minutes before the end of the incubation, 3 U of DNase RQ1 (Promega) were added. RNAs were then purified by phenol-chloroform extraction and ethanol precipitated prior to being dissolved in 30 µL of water. RNA products were then fractionated by denaturing (8 M urea) 7–10% (depending of the length of the candidates) polyacrylamide gel electrophoresis (PAGE ; 19:1 acrylamide : bisacrylamide) using a 45 mM Tris-borate, pH 7.5, 1 mM EDTA running buffer. RNA product bands were visualized by ultraviolet shadowing, and those of the correct sizes were excised from the gel and the transcripts eluted overnight at 4°C in elution buffer (1 mM EDTA, 0.1% SDS and 0.5 M ammonium acetate). RNA PG4s were then ethanol-precipitated, dried and dissolved in water. Concentrations were determined by spectrophotometry at 260 nm using a GE Nanovue spectrometer.

**Circular dichroism spectroscopy and thermal denaturation analysis**

Circular dichroism (CD) experiments were performed using 4 µM of the relevant RNA sample in 50 mM Tris-HCl (pH 7.5) buffer either in the absence of salt, or in the presence of 100 mM of either LiCl, NaCl or KCl. Before taking the CD measurement, each sample was heated at 70°C for 5 min and then slow-cooled to room temperature over a 1 h period. Experiments were performed using a Jasco J-810 spectropolarimeter equipped with a Jasco Peltier temperature controller in a 1-ml quartz cell cuvette with a pathlength of 1 mm. CD scans, ranging from 220 to 320 nm, were recorded at 25°C at a scanning speed of 50 nm $min^{-1}$ with a 2 s response time, 0.1 nm pitch and 1 nm bandwidth. All CD data represent the average of three wavelength scans. Subtraction of the buffer was not required as control experiments performed in the absence of RNA showed negligible curves. For thermal denaturation analysis, samples were heated from 25°C to 90°C at a controlled rate of 1° $min^{-1}$ and a 264 nm CD peak was monitored every 0.2 min in order to obtain the CD melting curves. Melting temperature values ($T_m$) were calculated using “fraction folded” (θ) versus temperature plots (Mergny et Lacroix, 2009).

**RNA labelling**

Before radioactive 5’-end-labelling, 50 pmol of purified RNA transcripts were dephosphorylated using 1 U of antarctic phosphatase (New England BioLabs) in a 10 µL final volume reaction containing 50 mM Bis-Tris-propane, pH 6.0, 1 mM $MgCl_2$, 0.1 mM

$ZnCl_2$ and 20 U RNase OUT (Invitrogen). Reactions were incubated at 37°C for 30 min and the enzyme was then inactivated by incubating at 65°C for 7 min. For the 5' end-labelling reaction itself, dephosphorylated transcripts (10 pmol) were incubated at 37°C for 1 h in the presence of 3 U of T4 polynucleotide kinase (USB), 3.2 pmol of [$\gamma$-$^{32}$P] ATP (6000 Ci/mmol ; New England Nuclear), 20 U of RNase OUT (Invitrogen) and in a buffer with final concentrations of 50 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 10 mM β-mercaptoethanol, all in a final volume of 10 µL. Labelling reactions were stopped by the addition of 10 µL of formamide dye buffer [95% formamide, 10 mM EDTA, 0,025% bromophenol blue (BPB) and 0,025% xylene cyanol (XC)]. Radiolabelled transcripts were fractionated by 7–10% denaturing polyacrylamide gel electrophoresis, and the bands were visualized by autoradiography. Those bands corresponding to transcripts of the correct sizes were excised from the gel and RNAs recovered and purified as described above. Purified 5'-end-labelled transcripts were dissolved in 30 µL of nanopure water, and radioactivity (total cpm) was measured using the Cerenkov method and Bioscan QC-2000 radioactivity counter.

**In-line probing**

Trace amounts (50 000 cpm, < 1 nM) of each 5'-end-labelled RNA were heated to 70°C for 5 min and then slow-cooled to 25ºC (≈ 1 h) in a buffer containing 20 mM lithium cacodylate (pH 7.5) and either in the absence of salt, or in the presence of 100 mM of either LiCl, NaCl or KCl (depending on the condition tested), all in a final volume of 10 µL. After slow-cooling, the volume of each reaction was adjusted to 100 µL in order to obtain final concentrations of 20 mM lithium cacodylate (pH 8.5), 20 mM $MgCl_2$ and 100 mM of LiCl, NaCl or KCl. Reactions were incubated for 40 h at 25ºC in order to allow for self-cleavage of the RNA to occur. After incubation, samples were ethanol-precipitated in presence of glycogen, ethanol-washed and dissolved in 10 µL of formamide dye buffer (that contained only XC as marker dye). An alkaline hydrolysis ladder was prepared with 50 000 cpm of the 5'-end-labelled transcript that was dissolved in 5 µL of water. One microliter of 1 N NaOH was added, and reactions were then incubated for 1 min at 25ºC. Reactions were stopped by addition of 3 µL of 1 M Tris-HCl pH 7.5, and then ethanol precipitated and dissolved in 10 µL of formamide loading dye (that contained only XC as marker dye). RNase T1 ladder was prepared by the addition of 0.6 U of RNase T1 (Roche Diagnostic) to 50 000 cpm of 5'-end-labelled transcript that was dissolved in 9 µL of buffer containing 20 mM Tris-HCl

pH 7.5, 10 mM $MgCl_2$ and 100 mM LiCl. Reactions were incubated for 2 min at 37°C, and then stopped by the addition of 20 µL of formamide dye buffer (that contained only XC as marker dye). Both samples and ladders were transferred to new eppendorf tubes and radioactivity (total cpm) measured using a Bioscan QC-2000 radioactivity counter. Both samples and alkaline hydrolysis ladder were then diluted in order to obtain equal amounts of radioactivity (cpm) for each loading sample, and ~two-third of these amounts of radioactivity for RNase T1 ladders. Samples and ladders were then fractionated by denaturing (8 M urea) 7–10% (depending on candidate size) polyacrylamide gel electrophoresis. Gels were dried and exposed overnight to a phosphoscreen. Bands were visualised by phosphorimaging using a Typhoon Trio instrument (GE Healthcare). Quantitative analyses of the bands were performed using the SAFA software (Das *et al.*, 2005). Two independent in-line probing experiments were performed and quantified for each candidate. Results are presented as one representative gel and a bar graph of the means and standard deviations of $K^+/Li^+$ intensities' ratios obtained from both experiments. A candidate is considered positive for G4 folding if the $K^+/Li^+$ ratio is $\geq 2$ (or significantly different from 1) for at least 2 nt predicted to be located in the loops and/or immediately next to the first or last G-tract. Banding pattern must differ from that of the mutated version abolishing G4 folding. Characteristics needed to be reproducible in at least two independent experiments.

**Plasmid constructions**

The sequences of the full-length 3'-UTRs of LRP5 and TTYH1 were obtained from the NCBI database and correspond to the following GenBank Accession numbers: LRP5, NM_002335 ; TTYH1, NM_020659. The 3'-UTRs were reconstituted *in vitro* by the filling in of multiple overlapping oligonucleotides and various PCR steps. The other plasmid constructs (TTYH1 + pAS, TTYH1 LRP5-pAS, LRP5 Ty PG4) were obtained by oligonucleotide-directed mutagenesis (see **Supplementary Tables S3 and S4 in Annexe 2** for both the detailed sequences and the list of the oligonucleotides used). Both the wild-type (WT) and a G/A mutant were synthesized for all 3'-UTR sequences. C/A- and GC/AA-mutant versions of the TTYH1 3'-UTR were also synthesized. The positions of the mutations were identical to those used for the *in vitro* in-line probing experiments. The different 3'-UTR constructs were inserted into the *Xba*I and *BamH*I restriction sites in the pGL3 plasmid

vector (Promega). The correct insertion of each construct was confirmed by DNA sequencing.

**Cell culture**

HEK293T cells were cultured in Dulbecco's Modified Eagle Medium supplemented with 10% fetal bovine serum, 1 mM sodium pyruvate and an antibiotic-antimycotic drug mixture (all from Wisent) at 37°C in a 5% $CO_2$ humidified incubator.

**Dual luciferase assays**

HEK293T cells were seeded in either 24- ($1.7 \times 10^5$) or 48-well plates ($8.5 \times 10^4$) 24 h prior to transfection. Cells were co-transfected with 400 ng of the specific pGL3 plasmid construct (Firefly luciferase, Fluc) and 100 ng of the pRL-TK control vector (Renilla luciferase, Rluc) (Promega) using Lipofectamine 2000 (Invitrogen) in Opti-MEM (Gibco) lacking the antibiotic-antimycotic mixture. Twenty-four hours after transfection, cells were lysed using passive lysis buffer (Promega). Fluc and Rluc activities were measured using the Dual-luciferase Reporter Assay kit (Promega) according to the manufacturer's protocol on a GloMax 20/20 Luminometer (Promega). For each condition, the Fluc value was normalised by dividing it by the Rluc value. Ratios of normalised WT version to normalised G/A-mutant version were calculated. Results are presented as means and standard deviations of a minimum of three independent experiments for each candidate.

## RESULTS AND DISCUSSION

**The PG4 sequence in the 3'UTR of TTYH1 folds *in vitro***

It has been previously demonstrated that G4s present in 3'-UTRs of mRNAs can stimulate polyadenylation when they are located downstream of a polyadenylation site (Beaudoin et Perreault, 2013 ; Decorsiere *et al.*, 2011). However, this has only been demonstrated in three distinct cases (i.e. LRP5, FXR1 and P53). In order to provide additional physical support for this phenomenon, the folding of the G4 motif within the TTYH1 gene (GenBank Accession number: GI 319803129, NM_020659) was investigated. The product of this gene is a calcium-independent, volume-sensitive chloride channel (Suzuki et Mizuno, 2004). This candidate was retrieved from a database that included all PG4s found in human 3'-UTR sequences using the classic algorithm search sequence $G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$, where

$x \geq 3$ and N is any nucleotide (A,C,G or U) (Beaudoin et Perreault, 2013 ; Huppert et Balasubramanian, 2005 ; Todd *et al.*, 2005). When considering the PG4 sequence alone, folding appeared highly probable (**Figure 19A**). The sequence bears 5 G-tracts and a few (1 to 7) intercalated nucleotides providing multiple possibilities of G4 conformations by various G-tract combinations. This was not possible in the three cases (LRP5, FXR1 and P53) reported previously. The TTYH1 PG4 candidate was also chosen because its 3'-UTR was relatively short (i.e. 348 nt), thereby circumventing a number of potential difficulties in the cloning step required for subsequent *in cellulo* investigations.

Initially, both G4 folding and topology were assessed by circular dichroism (CD) spectroscopy. A positive peak at 264 nm and a negative peak at 240 nm are characteristic of a parallel G4 topology (Paramasivan *et al.*, 2007). A WT sequence exceeding the TTYH1 PG4 in length by 12 nt in 5' and 13 nt in 3' was studied in order to assess G4-folding *in vitro* (see **Figure 19A**). Sequences at both extremities preserved some of the natural 3'-UTR genomic context (Arora *et al.*, 2009 ; Beaudoin et Perreault, 2013). A G/A-mutant version bearing six substitution adenines that is at least one in each G-tract which prevented G4 folding, was also *in vitro* transcribed for use as a negative control (see **Figure 19A**). Initially, CD spectrums were recorded either in the absence of salt, or in the presence of 100 mM of LiCl. No significant difference was observed between WT and G/A-mutant versions, confirming that the G4 folding did not occur under these conditions (see **Figure 19B**). The spectrum was then recorded in the presence of 100 mM of either NaCl or KCl (i.e. conditions that support G4 folding). Significant changes in both the 240 and 264 nm peaks were detected only for the WT sequence, confirming that G-quartets were stabilized by the monovalent cations, especially $K^+$ (**Figure 19B**). The conclusions drawn from the CD experiments received additional physical support from thermal denaturation analysis, where the melting temperature ($T_m$) for the WT version was found to be higher under the $K^+$ condition. In contrast, results for the G/A mutant showed no difference between the salt conditions (**Figure 19C**).

Further *in vitro* support was obtained from in-line probing of both WT and G/A-mutant PG4s. This simple technique which uses only trace amounts of radiolabeled RNA molecules (<1 nM), and thus favours intramolecular G4 folding, relies on the spontaneous cleavage of RNA under physiological conditions (Beaudoin *et al.*, 2013). Flexible regions,

such as single-stranded nucleotides, are relatively more prone to cleavage. In G4s, the connecting loops between G-tracts are typically flexible. TTYH1 transcripts were $^{32}$P-5'-labeled prior to being incubated for 40 h at 25ºC in the presence of $MgCl_2$ and either in the absence of salt, or in the presence of 100 mM of either LiCl, NaCl or KCl. Resulting samples were fractionated by denaturing (8 M urea) gel electrophoresis. Several WT transcript bands from the KCl sample (a condition that is a favourable for G4 formation) were relatively more intense than corresponding bands from the other samples (**Figure 19D**). Illustrating band intensity data by means of a bar graph displaying variations as $K^+/Li^+$ ratios showed that nucleotides exhibiting the highest susceptibility to hydrolysis were located between G-tracts and immediately 5' and 3' of the first and last G-tracts, respectively, a situation that is typical of G4 folding (see also the underlined residues in the TTYH1 PG4 sequence, **Figure 19A**) (Beaudoin *et al.*, 2013). Thus, three distinct and complementary methods show that the TTYH1 PG4 folds *in vitro* in the presence of $K^+$.

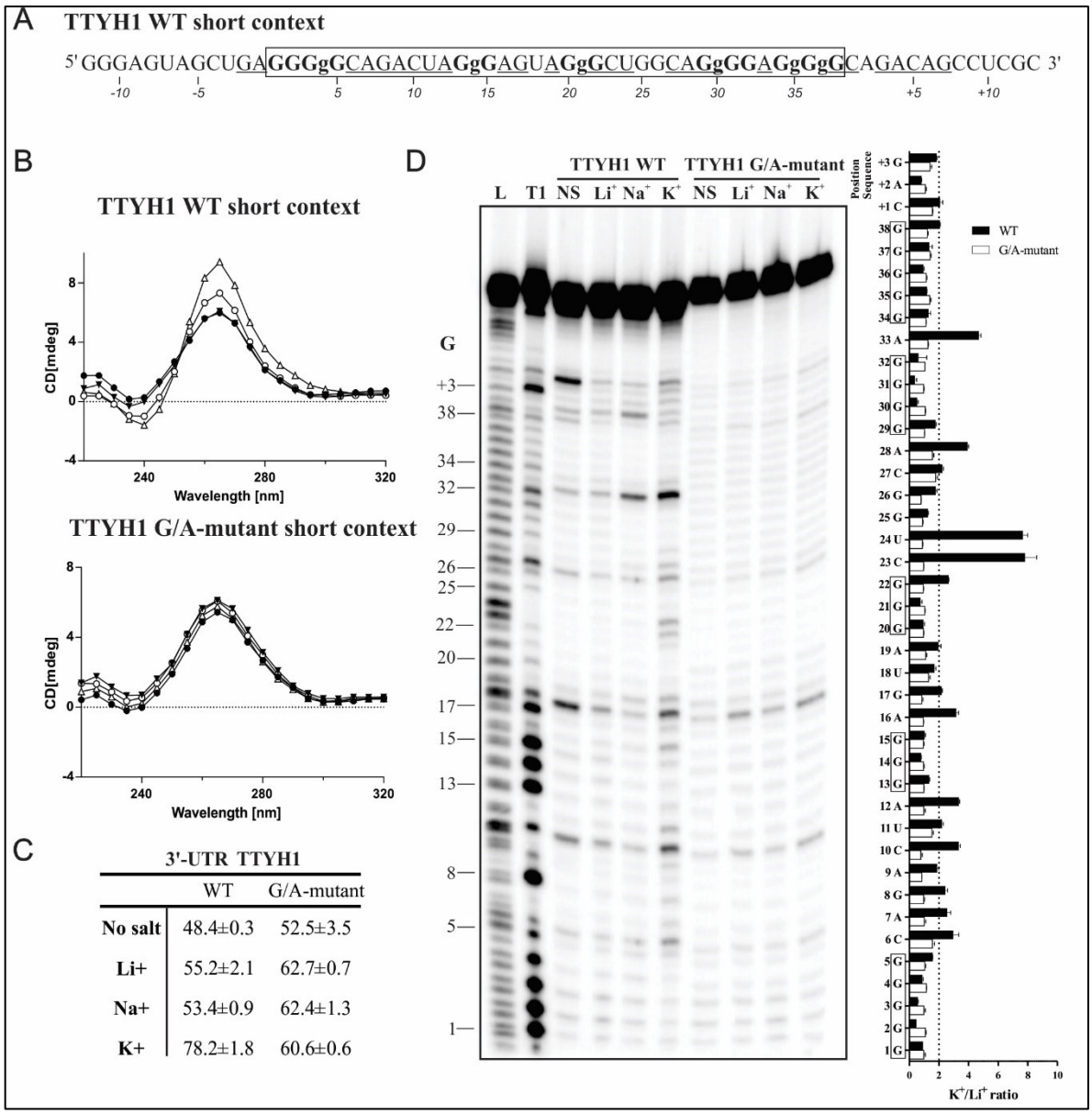



**Figure 19** – In vitro analysis of the TTYH1 WT PG4.

(A) Sequence of the TTYH1 WT PG4 surrounded by a short genomic context of 12-13 nt on both sides. The predicted PG4 sequence is composed of five G-tracts and is boxed. The guanines involved in the G-tracts are in bold. The lower case guanines (g) represent those that were mutated to adenine in the G/A-mutant version. Nucleotides showing greater cleavage accessibility in the in-line probing experiments are underlined. (B) CD spectroscopy analysis performed in absence of salt (filled circles), or in presence of 100 mM of either lithium ($Li^+$, filled triangles), sodium ($Na^+$, circles) or potassium ($K^+$, triangles). *Top panel:* The WT version shows a negative peak at 240 nm and a positive one at 264 nm which is characteristic of the formation of a G4 with a parallel topology. *Lower panel:* The equivalent spectrum for the G/A-mutant version in which some of the guanines of the G-tracts were mutated to adenine in order to abolish any possible G4 folding. (C) Thermal denaturation analysis of both the WT and the G/A-mutant versions in either absence of salt, or in the presence of 100 mM of either $Li^+$, $Na^+$ or $K^+$. Melting temperature values ($T_m$) were calculated using "fraction

folded" (θ) versus temperature plots (Mergny et Lacroix, 2009). (D) In-line probing analysis of both the TTYH1 WT and the G/A-mutant versions in either the absence of salt (No salt, NS), or in the presence of 100 mM of either $Li^+$, $Na^+$ or $K^+$. Bar graph represents the $K^+/Li^+$ intensity ratios of 2 independent experiments. Error bars represent standard deviation.

**Neighbouring C-rich sequences affect G4 folding**

The TTYH1 PG4 was further analysed in order to assess its ability to fold *in cellulo.* The full-length 3'-UTR sequence was cloned downstream of a firefly luciferase reporter gene and analysed using a dual luciferase system (see **Figure 20A** and the Material and Methods for details). HEK293T cells were co-transfected with plasmids expressing both the firefly and renilla luciferases (for normalization), grown for 24 h and then harvested for the performance of the luciferase assays. As 3'-UTR G4s are known to stimulate gene expression, a 2-fold difference in protein synthesis of the WT over G/A-mutant would suggest G4 folding (see **Figure 20B**). The 3'-UTR of LRP5 was used as a positive control and showed a ~2-fold difference, reflecting a higher level of protein synthesis during G4 folding [**Figure 20B *(i)***]. A previous report suggested that the 3'-UTR LRP5 PG4 folded into a G4 *in cellulo*, stimulating polyadenylation at a non-canonical upstream site (Beaudoin et Perreault, 2013). For the TTYH1 WT PG4, protein synthesis remained unchanged suggesting an absence of G4 folding in the context of the native full-length sequence [**Figure 20B *(ii)***]. The TTYH1 PG4 sequence commences at the 120th nt position of the 3'-UTR and bears a canonical poly-adenylation signal (pAS) commencing at the 168th nt position downstream of the g-quadruplex. The absence of any significative difference in luciferase expression for the TTYH1 WT construct may result from the fact that TTYH1 lacks an upstream polyadenylation signal as opposed to LRP5 (see the comparative schematics of the 3'-UTR architectures under **Figures 20B *(i)* and *(ii)***). In order to verify this hypothesis, a polyadenylation signal was inserted 49 nts upstream of the TTYH1 PG4 that is in a position analogous to its location in the LRP5 sequence [TTYH1 + pAS, **Figure 20 *(iii)***]. In addition, the canonical signal was mutated in order to force polyadenylation to occur at a position potentially stimulated by the PG4, once again to mimic the effect of the LRP5 G4. No significant difference was observed with this construct [**Figure 20B *(iii)***], suggesting that either the TTYH1 PG4 remained unfolded *in cellulo*, or that a required co-factor was lacking. The LRP5 PG4 sequence was then substituted for the TTYH1 PG4 sequence while conserving the remaining LRP5 3'-UTR intact [i.e. LRP5 Ty-PG4; **Figure 20B *(iv)***]. This

construct displayed a significant 3-fold increase in protein synthesis compared to the LRP5 WT, indicating that the TTYH1 PG4 folded *in cellulo* and stimulated polyadenylation more efficiently than the LRP5 G4.

Results presented above suggested that unaltered protein synthesis for TTYH1 PG4 within its natural 3'-UTR context may be attributable to the composition of the neighbouring sequences. To further investigate this possibility, several constructs were engineered. A first construct was synthesized by inserting not only the 6 nts of the LRP5 polyadenylation signal of the TTYH1 + pAS version, but also additional nucleotides located between this polyadenylation signal and the first G-tract of the LRP5 PG4, creating the TTYH1-LRP5-pAS construct [**Figure 20B *(v)***]. Thus this construct would contain potential *cis*-regulating elements important for polyadenylation, such as the U-rich region located downstream of the cleavage site, as well as both the primary and secondary structures surrounding it. However, still no significant difference in protein synthesis was observed [**Figure 20B *(v)***]. Alternatively, we hypothesized that the absence of effect might be due to folding hindrance attributable to C-tracts located (13–50 nt) mainly downstream but also upstream of the PG4. An earlier study revealing a large number of neighbouring cytosine residues in a PG4 retrieved in the 5'-UTR region of three mRNAs impaired for G4 folding (Beaudoin et Perreault, 2010), suggested that C-rich regions could compete with G-tracts to form Watson-Crick base pairs, thereby hindering G4 folding. Neighbouring cytosine residues were, however, not part of the short transcript studied previously *in vitro,* or the LRP5 Ty-PG4 construct. We then replaced certain cytosine residues by adenines in the TTYH1 and TTYH1 + pAS C/A-mutants [**Figure 20B *(vi)* and *(vii)***, respectively]. Still no effect was observed with either of these C/A-mutants. Importantly, both constructs were deprived of essential *cis*-regulatory elements. A polyadenylation signal was also absent from the TTYH1 C/A-mutant. Next, we engineered a TTYH1-LRP5-pAS C/A-mutant version including both *cis*-regulatory elements and a polyadenylation signal [**Figure 20B *(viii)***]. This construct exhibited a significant 2-fold increase in luciferase expression. Taken together, these data suggest that a series of cytosines located as far as 20–50 nucleotides from the PG4 can significantly influence G4 folding *in cellulo*, demonstrating that regulation of G4 folding is far more complex than what has been previously reported (Beaudoin et Perreault, 2010, 2013).

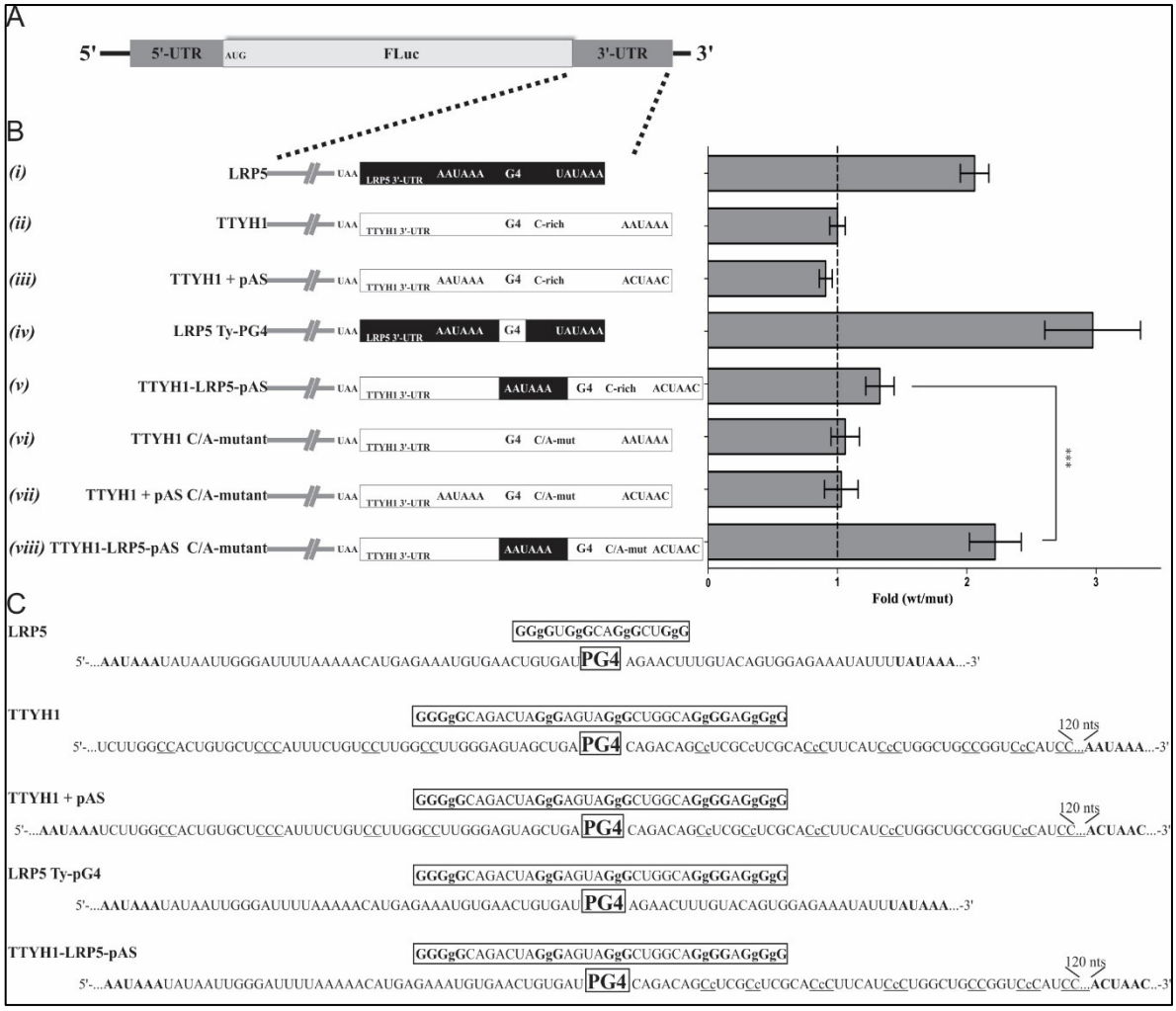



**Figure 20** – Luciferase assays measuring the effects of the 3'-UTR G4s on the stimulation of gene expression.

(A) Schematic representation of the firefly luciferase reporter gene construction transfected into HEK293T cells. The full sequences of all of the constructs used are listed in **Supplementary Table S3 in Annexe 2** (B) *Left panel:* Schematic representation of all of the full-length 3'-UTRs used. The black and white rectangles represent the sequences of LRP5 and TTYH1, respectively. The constructs with 'pAS' had their canonical polyadenylation site (AAUAAA) abolish through mutations to ACUAAC. Those identified as C/A-mutant had several of cytosines located downstream of the G4 mutated to adenines to abolish the C-tracts. *Right panel:* Gene expression levels of the different constructs as measured at the protein level with the firefly luciferase assay, are represented as a fold difference of the value obtained for the G4 WT version divided by that obtained for the G4 mutant version (in which key guanines of the G4 were mutated to adenines to abolish G4 folding). Error bars (standard deviations) were calculated with the results of at least three independent experiments. *P*-values were evaluated with a two-tailed, unpaired Student t-test. ***P<0.0001 (C) Partial sequences of the different 3'-UTR constructs. The PG4 region is boxed, the guanines of the G-tracts and the polyadenylation signal (pAS) are shown in bold and the C-rich regions are underlined. The lowercase cytosines (c) are those mutated to adenines in the C/A-mutant versions.

**In-line probing of TTYH1-derived transcripts**

To further investigate the influence of C-rich regions on TTYH1 G4 folding, *in vitro* in-line probing experiments were performed. A short and a long transcript corresponding to the TTYH1-LRP5-pAS WT were synthesized. A TTYH1-LRP5-pAS C/A-mutant was synthesized in the long version only. In the short version, the PG4 was flanked in 5' and 3' by 9 and 13 nt respectively. In the long version, the PG4 was flanked on each side by a 50-nt sequence found in the mRNA (**Figure 21A**). For all transcripts, a G-tract G/A-mutant version was also engineered for purposes of comparison between the presence and absence of PG4. An autoradiogram of a representative in-line probing for the TTYH1-LRP5-pAS transcript is illustrated on the left panel of **Figure 21B**. The right panel of **Figure 21B** shows the quantitative analysis of two independent experiments in the form of a bar graph representing $K^+/Li^+$ intensity ratios. Nucleotides with a ratio value superior to the arbitrary threshold of 2 were considered as being accessible under conditions favouring G4 folding (Beaudoin *et al.*, 2013). Band intensities increased for WT version residues located in single-stranded regions adjacent to G-tracts on G4 folding and solely in the presence of $K^+$. These residues correspond to nucleotides $A_{16}$, $C_{23,}$ $U_{24}$ and $A_{33}$. While the TTYH1-LRP5-pAS short transcript folded into a G4, the same PG4 located within the long transcripts displayed no significant difference in banding patterns compared to WT and G/A-mutant versions, regardless of whether incubation was performed in the presence of LiCl or KCl (**Figure 21C**). These results indicated that the PG4 did not fold differently in the presence of $K^+$. Conversely, the C/A-mutant version bearing five substitute adenosines between positions 9–43 displayed two additional bands of greater intensity in the presence of KCl, compared to both WT and G/A-mutant versions (**Figure 21C**). Both of these corresponded to nucleotides $C_{23}$ and $A_{33}$, as was the case for the short version. These data show that mutating the C-tract to hinder potential GC Watson-Crick base-pairs was sufficient to favour G4 folding. Taken together, these results confirm that C-tracts located up to 20–50 nts from the PG4 can prevent G4 folding, and that folding can be rescued where C-tracts are either completely absent as in the short context version, or replaced by adenines as in the C/A-mutant version.

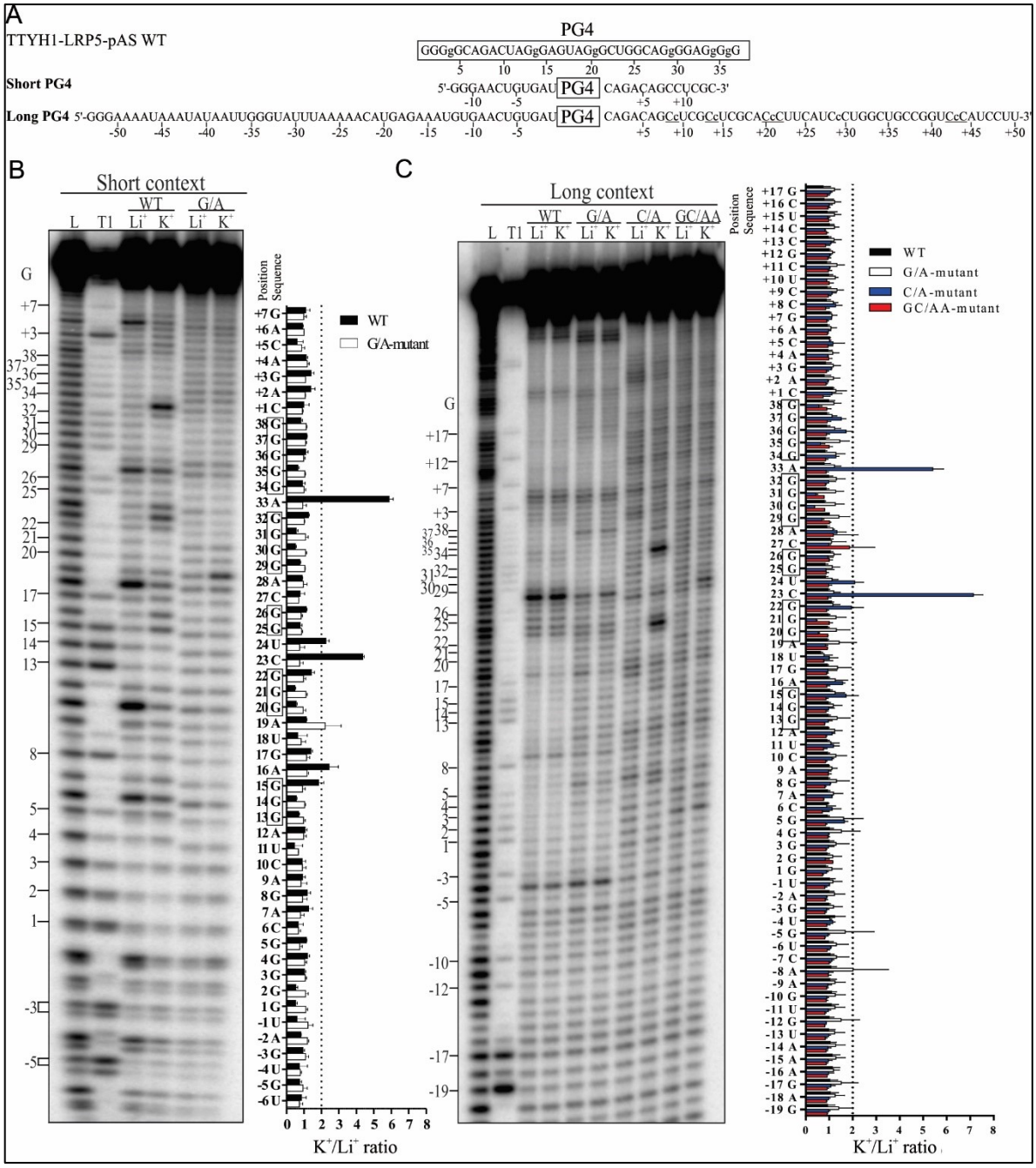



**Figure 21** – In-line probing and quantitative analysis of structures adopted by the TTYH1-LRP5-pAS PG4 candidate in both short and long genomic contexts.

(A) Sequences of the TTYH1-LRP5-pAS PG4 candidate with its short (~13 nt each side) and long (~50 nt each side) genomic contexts. The predicted PG4 sequence is boxed. The guanines of the G-tracts are shown in bold. Nucleotide positions are indicated under the sequence and '+' and '-' indicate whether the genomic nucleotide is located downstream or upstream of the PG4, respectively. The lowercase guanines (g) and cytosines (c) represent those mutated to adenines in the G/A- and C/A-mutant versions, respectively. The C-tracts are underlined. (B) Autoradiogram of a 10% denaturing (8 M urea) polyacrylamide gel of the in-line probing of both the WT and G/A-mutant

versions of the TTYH1-LRP5-pAS in their short genomic contexts. (C) Autoradiogram of a 7% denaturing (8 M urea) polyacrylamide gel of the in-line probing of the WT, G/A-, C/A-and GC/AA-mutant versions of the TTYH1-LRP5-pAS in the long genomic context. For both genomic context lengths (B and C) the in-line probings were performed in presence of 100 mM of either LiCl ($Li^+$) or KCl ($K^+$). Positions of the guanines are indicated on the left of each gel. L and T1 are the alkaline hydrolysis and ribonuclease T1 mapping lanes, respectively. The bar graphs (right portions of panels B and C) represent the quantitative analysis of the in-line probing and show, for each band, the intensity values in the $K^+$ condition divided by that in the $Li^+$ condition. The dotted line corresponds to the threshold of two that denotes the nucleotides with higher accessibilities to in-line cleavage in the presence of $K^+$ (i.e. the nucleotides located in the G4 loop). The G-tracts predicted to be implicated in the G4 formation are boxed. Black, white, blue and red represent the WT, G/A-mutant, C/A-mutant and GC/AA-mutant versions, respectively. Error bars are standard deviations as calculated from two independent experiments. The G4 is formed only in the cases of the short context WT and the long C/A-mutant version context.

**Assessing the PG4 genomic context**

Results with the TTYH1-LRP5-pAS transcript unambiguously showed that PG4 neighbouring sequences have a significant impact on G4 folding. Next, to broaden the scope of this study, we investigated 11 other PG4s from human 5'- and 3'-UTRs, as well as three C/A-mutant versions of 5'-UTR PG4s for which mutations were shown to be required for G4 folding (Beaudoin et Perreault, 2010). All of these candidates were previously characterized by in-line probing as part of short transcripts bearing ~15 nt in 5' and 3' of the PG4 (Beaudoin et Perreault, 2010 ; Biffi *et al.*, 2013). Five variants of the TTYH1-derived constructs were also considered, for 19 candidates. In-line probing was repeated using larger transcripts in which both WT and G/A-mutant versions of PG4s were flanked by ~50 nts on both sides. The resulting bar graphs for each candidate are presented in supplementary **Figures S1–S12 in Annexe 2**, and a data compilation is given in **Table 3**. Shaded candidates are those that do not fold into G4s. Thirteen out of the nineteen candidates supported G4 folding of the longer transcripts in the presence of KCl. It is noteworthy that results obtained with both short and long transcripts are in agreement for most of the candidates. In other words, PG4s folded regardless of the size of neighbouring sequences. However, this was not so for four of the candidates that are the DOC2B and TNFSF12 C/A-mutants, and the TTYH1 and MAPK3 WTs (**Table 3**). Further analysis of the primary sequence of each candidate provided an interesting explanation (**Figure 22A-D**). The short transcripts of the DOC2B and TNFSF12 C/A-mutants and the TTYH1 WT all support G4 folding in the presence of KCl, but their longer counterparts do not. The extended sequences of these candidates included several C-tracts that most likely form GC Watson-Crick base pairs with residues of the G-tracts, thereby

preventing G4 folding (see **Figure 22A–C**). Some of the possible inhibitory cytosines located in close proximity to the PG4 were replaced by adenines in both the DOC2B and TNFSF12 C/A-mutants. However, cytosines located farther than 15 nt away from the PG4 can still base-pair with the guanines from the G-tracts. The MAP3K11 PG4 WT candidate is slightly different, but respects the previous logical assumption (**Figure 22D**). For this candidate, the short transcript was unable to fold into a G4, most likely due to C-tracts adjacent to the PG4 competing to form inhibitory Watson-Crick base pairs. The longer transcript however bore several additional G-tracts which probably interacted with C-tract residues, thereby releasing PG4 and allowing it to fold into a G4 in the presence of KCl. This hypothesis is supported by RNAfold (Hofacker *et al.*, 1994) secondary structure predictions for both short and long transcripts (**Supplementary Figure S27 in Annexe 2**). Although for most PG4s studied, the length *per se* of neighbouring sequences did not impact G4 folding, results for four of the candidates point to the inherent complexity of neighbouring sequences as a key issue that must be considered in the accurate prediction of biologically relevant G4 motifs.

**Table 3** Characteristics of selected PG4 candidates

| | Long context | | | | | Short context | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Candidates** | **Total Loop Length** | **Mfe** | **cG score** | **cC score** | **cG/cC** | **Total Loop Length** | **Mfe** | **cG score** | **cC score** | **cG/cC** | **Agreement** |
| **NCAM2 WT** | 14 | -54.9 | 1470 | 370 | 4.0 | 14 | -20.4 | 840 | 170 | 4.9 | + |
| **BARHL1 WT** | 6 | -55.5 | 1870 | 720 | 2.6 | 6 | -8.5 | 1350 | 50 | 27 | + |
| **FZD2 WT** | 6 | -54.1 | 3420 | 450 | 7.6 | 6 | -16.8 | 2400 | 180 | 13.3 | + |
| **EBAG9 WT** | 4 | -49.2 | 1410 | 650 | 2.2 | 4 | -13.9 | 1060 | 150 | 7.1 | + |
| **FXR1 WT** | 7 | -29.3 | 1330 | 200 | 6.7 | 6 | -8.2 | 1170 | 130 | 9 | + |
| **LRP5 WT** | 5 | -15.7 | 860 | 80 | 10.8 | 5 | -6.5 | 610 | 50 | 12.2 | + |
| **AASDHPPT WT** | 4 | -45.6 | 1190 | 570 | 2.1 | 4 | -13.7 | 820 | 100 | 8.2 | + |
| **THRA1 WT** | 11 | -54.1 | 2280 | 670 | 3.4 | 11 | -19.7 | 970 | 200 | 4.9 | + |
| **DOC2B WT** | 12 | -71.1 | 1370 | 1690 | 0.8 | 12 | -31.6 | 880 | 730 | 1.2 | + |
| **DOC2B C/A-mutant** | 12 | -55.2 | 1370 | 1170 | 1.2 | 12 | -10.8 | 880 | 210 | 4.2 | - |
| **TNFSF12 WT** | 7 | -53.8 | 1450 | 2510 | 0.6 | 7 | -30.4 | 1220 | 940 | 1.3 | + |
| **TNFSF12 C/A-mutant** | 7 | -45.9 | 1450 | 1740 | 0.8 | 7 | -10.3 | 1220 | 170 | 7.2 | - |
| **MAP3K11 WT** | 11 | -69.8 | 2160 | 760 | 2.8 | 11 | -23.9 | 710 | 480 | 1.5 | - |
| **MAP3K11 C/A-mutant** | 11 | -50.1 | 2160 | 440 | 4,9 | 11 | -12.3 | 710 | 160 | 4.4 | + |
| **TTYH1 WT** | 12 | -62.1 | 1580 | 780 | 2.0 | 12 | -15.0 | 1320 | 130 | 10.2 | - |
| **TTYH1 C/A-1-mutant** | 12 | -48.3 | 1580 | 480 | 3.3 | | | | | | |
| **TTYH1 C/A-2-mutant** | 12 | -34.4 | 1580 | 350 | 4.5 | | | | | | |
| **TTYH1-LRP5-pAS WT** | 12 | -54.2 | 1550 | 590 | 2.6 | 12 | -13.4 | 1310 | 130 | 10.1 | - |
| **TTYH1-LRP5-pAS C/A-mutant** | 12 | -29.8 | 1550 | 290 | 5.3 | | | | | | |

Shaded entries indicate those candidates that cannot fold into G4.

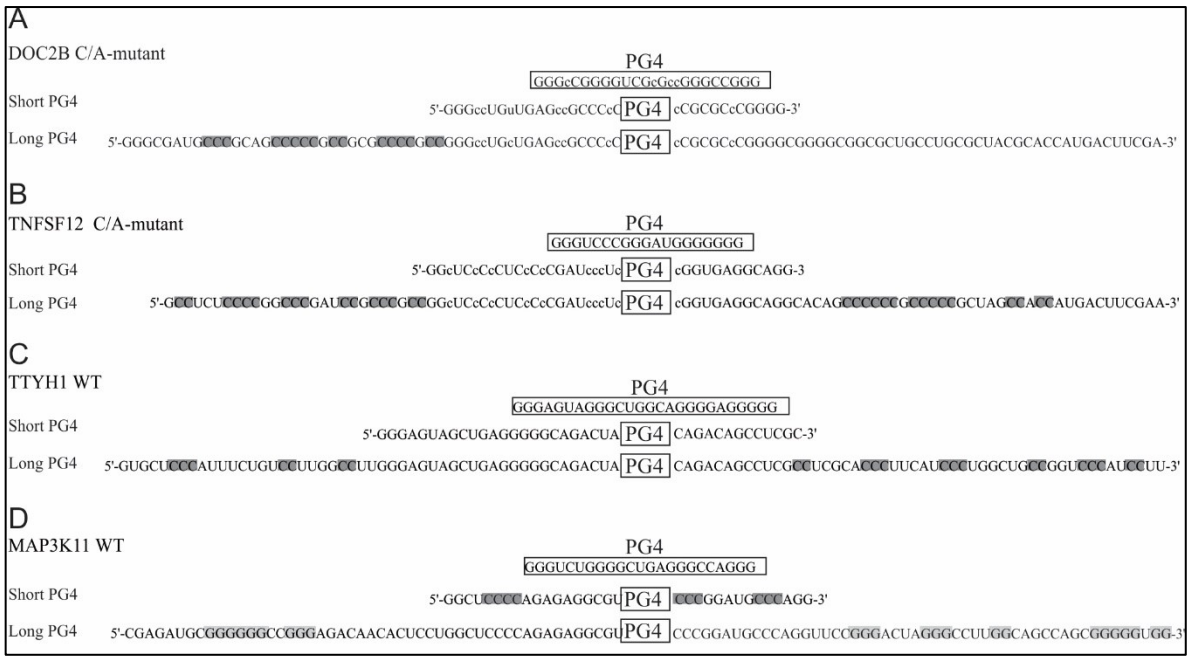



**Figure 22** – Sequence analysis of the genomic context of non-folding PG4s.

(A) DOC2B C/A-mutant ; (B) TNFSF12 C/A-mutant ; (C) TTYH1 WT; and, (D) MAP3K11 WT. The PG4 sequences predicted by the algorithm are boxed. Guanines involved in the G-tracts are shown in bold. Sequences of both the short and the long genomic context versions are presented. Lower case cytosines (c) represent those mutated to adenines in the C/A-mutants. Inhibitory tracts of cytosines present in the genomic context version are highlighted in dark gray. Enhancing tracts of guanines are highlighted in pale gray.

### Determining a predictive parameter of G4 folding

Next, we sought to identify a reliable G4-folding predictive value tested against the data obtained for our set of characterized PG4s. Such a value should be able to indicate whether the genomic environment of a given PG4 is favourable or not to G4 folding. One of the most frequently used criteria for evaluating PG4 stability, and thus G4-folding probability, is the total loop length. For both DNA and RNA G4s, longer loops are associated with relatively less stable free energies (ΔG°) and melting temperatures (Tm) (Bugaut et Balasubramanian, 2008 ; Guédin *et al.*, 2010 ; Zhang *et al.*, 2011a). Accordingly, the lower the total loop length, the more likely the G4 folding. The total loop length for a given PG4 was simply calculated as the sum of the nucleotides present in each of its three loops (**Table 3**). Surprisingly, for the set of PG4s characterized in this study, total loop length was not a relevant indicator of G4 folding (**Figure 23A**). No significant difference in total loop length was observed between the folding and non-folding sequences. That said, the majority of PG4s with lower predicted total loop length folded into G4s. However, so did many of the PG4s with higher

total loop length e.g. NCAM2 WT, TTYH1 C/A-1 and -2 mutants. Furthermore, TNFSF12 WT and C/A-mutant sequences which had relatively lower total loop length did not fold into G4s. These results are in agreement with those of a previous study performed with human DNA promoter G4s showing that G4 stability did not correlate with loop length (Kumar et Maiti, 2008). For all of these reasons, total loop length did not appear to be a suitable predictive parameter of G4 folding.

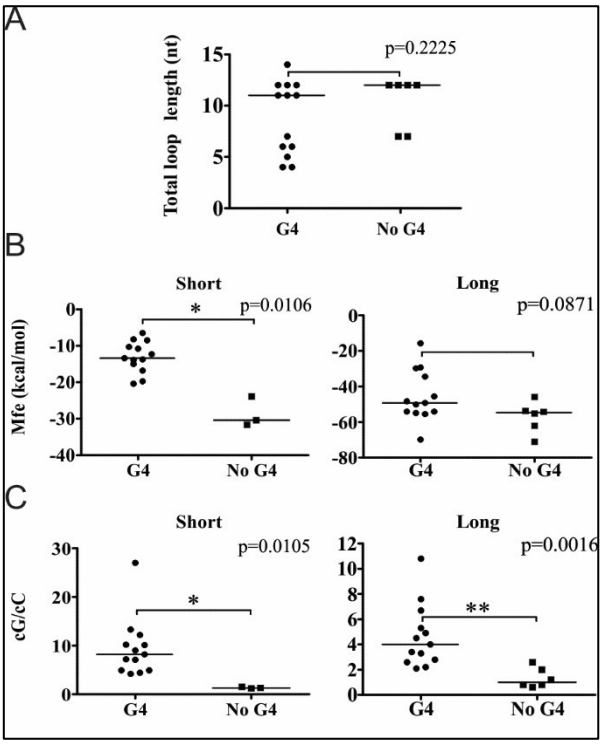



**Figure 23** – Comparison of the different predictive values of G4 folding for both short and long genomic context PG4 candidates.

(A) Predicted total loop length; (B) Mfe, kcal/mol, predicted by the RNAfold software from the Vienna RNA Package (Hofacker *et al.*, 1994); and, (C) calculated cG/cC score. Horizontal lines represent the median for each group. *P*-values were evaluated with a Mann-Whitney test. *P<0.05 **P<0.01.

As pointed out both here, and in previous reports (Arora *et al.*, 2009 ; Beaudoin et Perreault, 2010 ; Bugaut *et al.*, 2012), the genomic context of a PG4 may influence its folding. It is reasonable to hypothesize that if the genomic context of a given PG4 makes it prone to multiple and strong Watson-Crick base-pair based secondary structures, that this could hinder G4 folding. The thermodynamic stability of an RNA secondary structure can be conveniently estimated using prediction software such as RNAfold from the Vienna RNA Package

(Hofacker *et al.*, 1994). The software version used however considers only Watson-Crick and wobble base-pair formation, and therefore could not predict G4 folding. This software provides a Mfe for each predicted structure. According to the formulated hypothesis, a structure with a lower predicted Mfe would be less favorable to G4 folding because the stable Watson-Crick based secondary structure should form faster than the G4 motif. The secondary structures of both short and long transcripts were predicted using RNAfold, and the Mfe values for the most stable structures were compiled and analyzed together (see **Table 3 and Figure 23B**). For short transcripts, this value was an excellent indicator of G4-folding and non-folding ($P$= 0.0106). However, no significant differences were observed between the Mfe values for both folding and non-folding longer transcripts ($P$= 0.0871). It seems that when considering the longer transcripts, the possibilities of multiple secondary structures increases substantially. Consequently, the use of the Mfe as a predictor of G4 folding within any given transcript is not always valid.

In the absence of a suitable predictive parameter of G4 folding for long RNA transcripts, we attempted to identify one. By definition, G4 folding requires multiple G-tracts in a given sequence. However, there are recent examples of G4 with discontinuous G-tracts or with G-tracts bearing only two consecutives Gs (Mukundan et Phan, 2013 ; Mullen *et al.*, 2010). Nonetheless, these guanines must be primarily single-stranded, or otherwise sufficiently available, to interact with each other to fold into a G4. Conversely, consecutive cytosines (Cs) have been shown to potentially impair G4 folding, most likely due to pairing with consecutives Gs within a stable Watson-Crick base-paired structure. Following this rationale, separate consecutive G (cG) and consecutive C (cC) scores were considered (see Materials and Methods). Briefly, a score of 10 was attributed for each single G or C, a score of 20 for each doublet GG or CC, 30 for each triplet GGG or CCC and so on. We assumed that longer G-tracts should favor G4 folding, whereas longer C-tracts should hinder it. cG and cC scores were the respective sums of all values attributes to Gs and Cs for a given sequence. Thus, the longer the consecutive nucleotide tracts, the higher their score value. For example, a series of three consecutive Gs will a have a cG score of 100[($cG\ score = 3\ (G) \times 10 + 2(GG) \times 20 + 1(GGG) \times 30 = 100$], while a series of two consecutive Gs a score of 40 [$cG\ score = 2\ (G) \times 10 + 1(GG) \times 20 = 40$]. The cG and cC scores of all study candidates, in both the short and long contexts, were determined (**Table 3**). As expected,

analysis showed that both scores were higher for the longer sequences. Next, to define a parameter integrating both components, the cG score was divided by the cC score, providing the cG/cC score (see **Table 3 and Figure 23C**). The cG/cC score clustered the G4-folding and non-folding RNA species regardless of transcript length. *P*-values of 0.0105 and 0.0016 were estimated for the short and the long transcripts respectively. For short transcripts, the *P*-value (0.0105) is slightly lower than that obtained based on the Mfe (0.0106). Our cG/cC score is a novel predictor of RNA G4-folding which appears to significantly discriminate between folding and non-folding PG4s among a set of different RNA molecules.

**Challenging the cG/cC score**

To assess the predictive potential of the cG/cC score with respect to G4 folding, 14 novel PG4 candidates retrieved in human 5'-UTRs were selected for analysis. The candidates were chosen for the diversity of their genomic contexts, as illustrated by their predicted Mfe values ranging from -71.7 to -20.6 kcal/mol (**Table 4**). Both the WT and G/A-mutant versions of each PG4, flanked by ~50 nts on each side, were synthesized by run-off transcription, 5'-radiolabelled and then submitted to the in-line probing procedure described previously in the presence of 100 mM of either LiCl or KCl. The sequences of each PG4 are given in **Supplementary Table S2 in Annexe 2**, and the resulting bar graphs from the in-line probing experiments are presented in **Supplementary Figures S13 to S26 in Annexe 2**. Only six transcripts supported G4 folding in the presence of $K^+$ according to the in-line probing data (**Table 3**). The other eight transcripts are considered false positive predictions of the standard sequence algorithm. A bar graph displays the candidates in order of increasing predicted Mfe values, in **Figure 24C**. As expected, the four candidates with the highest predicted Mfes, that is those corresponding to relatively less stable structures, supported G4 folding whereas the three with the lowest Mfes did not permit G4 folding. However, the seven other candidates with middle Mfe values provided somewhat unexpected results. Mfes for the cluster of G4-folding G4 candidates versus that for the non-folding candidates provide a *P*-value of only 0.0813 (**Figure 24A**), showing no significant difference between both groups. This is further evidence that Mfe is not an accurate predictive parameter of G4 folding for RNA molecules bearing flanking sequences that mimic the arrangement in naturally-occurring transcripts. No total loop length difference was observed between folding and non-folding candidates, indicating that this parameter is also ineffective predictor of G4 folding (data not shown).

**Table 4** Characteristics of PG4 candidates selected to challenge the predictive parameters.

| Candidates | Total Loop Length | Mfe | cG score | cC score | cG/cC |
|---|---|---|---|---|---|
| **MAPK3 WT** | 11 | -20.6 | 1890 | 200 | 9.5 |
| **SYNCRIP WT** | 15 | -30.6 | 2850 | 240 | 11.9 |
| **PPP1CA WT** | 15 | -34.2 | 1650 | 280 | 5.9 |
| **ERCC2 WT** | 12 | -38.2 | 2700 | 430 | 6.3 |
| **ESR2 WT** | 11 | -46.8 | 1140 | 1610 | 0.7 |
| **TCF7L1 WT** | 12 | -47.9 | 1060 | 800 | 1.3 |
| **SMAD2 WT** | 14 | -52.7 | 1350 | 790 | 1.7 |
| **SMAD7 WT** | 12 | -53.7 | 1190 | 1500 | 0.8 |
| **DNMT3B WT** | 14 | -55.7 | 1620 | 540 | 3.0 |
| **CREM WT** | 11 | -56.2 | 1930 | 800 | 2.4 |
| **ACVR1C WT** | 15 | -63.3 | 2300 | 730 | 3.2 |
| **GNAI2 WT** | 13 | -68.4 | 2000 | 630 | 3.2 |
| **PTPRJ WT** | 14 | -69.8 | 1620 | 620 | 2.6 |
| **MYCL1 WT** | 12 | -71.7 | 1410 | 720 | 2.0 |

Shaded entries indicate those candidates that cannot fold into G4.

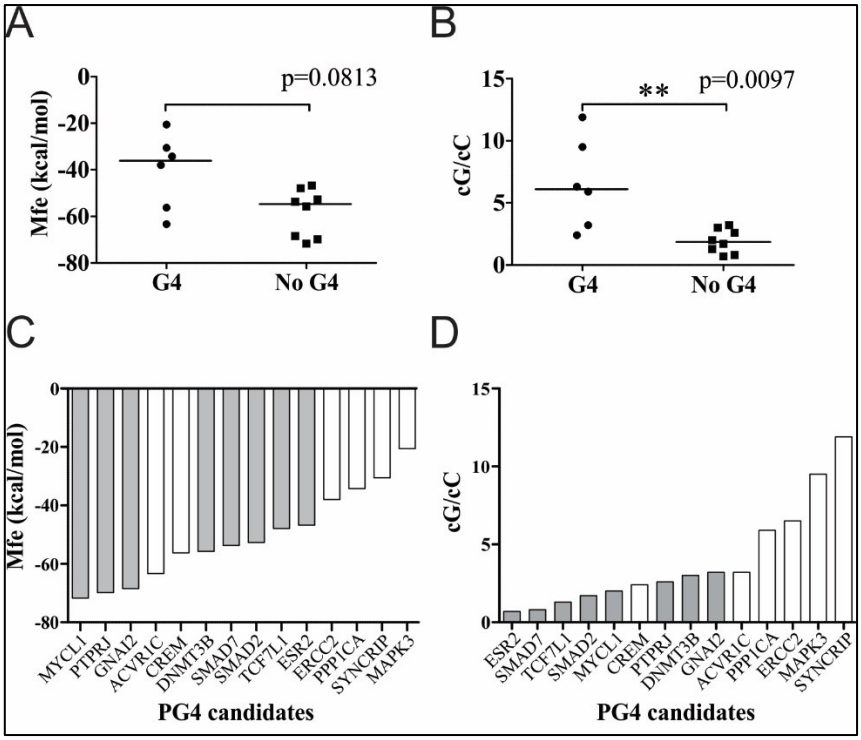



**Figure 24** – Challenge of predictive values for a new set of 14 PG4s with various genomic contexts.

(A) Mfe, kcal/mol ; and, (B) calculated cG/cC score. Horizontal lines represent the median for each group. *P*-values were determined with a Mann-Whitney test. **P<0.01. (C-D) Bar graphs representing the predicted Mfe value (C) and the cG/cC score (D) of the PG4 candidates placed in increasing order. White bars represent folding candidates, whereas gray bar represent non-folding candidates.

Next, the cG/cC scores were determined for all 14 candidates. Candidates were classified accordingly, the higher the cG/cC score, the more likely G4 folding and *vice versa*. For 13 out of 14 candidates, the scoring system was a strongly accurate predictor of G4 folding as illustrated in **Figure 24D**. The single exception, the CREM PG4, displayed an intermediate cG/cC score of 2.4. All other candidates with a high cG/cC score supported G4 folding, whereas low scorers were non-folding. When representing the cG/cC scores of folding and non-folding clusters of PG4 candidates, a *P*-value of 0.0097 indicated discrimination between both groups (**Figure 24B**). Results thus confirmed the predictive potential of the cG/cC score with respect to G4 folding of RNA transcripts in their genomic contexts. Moreover, the cG/cC score also seemed to limit the number of false positives predictions. Candidates with the lowest cG/cC scores can readily be regarded as non-folding PG4s.

**Comparison of the cG/cC score with existing parameters and predictive tools**

This study unambiguously demonstrated that, in contrast to the identification of the RNA PG4s, the accurate prediction of G4 folding requires careful consideration of both upstream and downstream sequences beyond just a few nucleotides on either side of the PG4. More or less distant C-tracts have been suggested to impair G4 folding. While both a conventional secondary structure prediction algorithm and the resulting Mfe parameter appear to be of limited use in determining whether a PG4 located in a relatively long RNA species will fold into a G4, the proposed cG/cC score appears to be a useful alternative. The ratio of cGs to cCs appears to be a good predictor of G4 folding for RNA transcripts. A relatively common way to assess the accuracy of a predictive test and to set a threshold between two conditions (folding and non-folding in the case at hand) is to draw ROC curves. In this kind of analysis, sensitivity – that is the fraction of G4 folding candidates with cG/cC scores above the threshold, is plotted against specificity – that is the fraction of non-folding candidates with scores below the threshold. The quality of the predictive value is the AUC. An area of 0.5 represents a random (i.e. non-discriminating value), while an AUC of 1 represents a perfect prediction (i.e. generating no false positives or negatives). ROC curves analyses demonstrated that the cG/cC score is both the most sensitive and specific predictor of G4 folding for long RNA transcripts. The cG/cC score also displays higher AUCs compared to total loop length and predicted Mfe (**Figure 25**, and the AUCs presented in **Supplementary Table S5 in Annexe 2**).

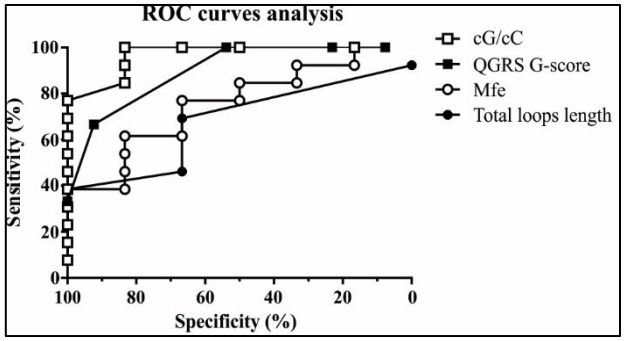



**Figure 25** – ROC curves analysis of the different predictive parameters for PG4 in their long context.

QGRS G-score was evaluated with the QGRS software (Kikin *et al.*, 2006). The AUC is the ability to discriminate between folding and non-folding PG4s. An AUC of 0.5 is a random discriminating value, while an AUC of 1 stands for perfect discrimination. The cG/cC score displays the highest

AUC and, thus, the highest sensitivity and specificity. ROC curves analyses were performed using GraphPad Prism version 5.02 for Windows (GraphPad Software, San Diego California USA).

Use of a scoring system to predict G4 folding has already been proposed by others. Previously, Kikin and collaborators developed the so-called G-score which was used in combination with the standard sequence algorithm in their QGRS Mapper tool to predict G4 folding (Kikin *et al.*, 2006). The G-score takes into account the number of Gs in G-tracts and the number and arrangement of loop nucleotides. PG4s with longer G-tracts and shorter loop lengths with evenly distributed nucleotides have higher G-scores and are more likely to fold into G4s. However, when applied to the set of RNA PG4 candidates used here, the G-score was less sensitive and specific than the cG/cC score (**Figure 25**). Since the G-score is mostly based on loop length, this could explain its poorer predictive value compared to the cG/cC score. Despite what was expected from the conclusions of previous studies (Guédin *et al.*, 2010; Zhang *et al.*, 2011a), the total loop length was not a significant predictor of G4 folding for the RNA transcripts used in this study. Thus, contrary to their DNA counterparts, RNA PG4 stability and folding potential seem to be relatively less sensitive to loop length and arrangement. However, the neighbouring genomic context and possible competing Watson-Crick structures seem to bear relatively more importance for predicting the folding potential of RNA PG4s. Its single-stranded nature affords RNA PG4s great plasticity, enabling the same molecule to rapidly adopt a multitude of stable secondary structures, whereas in DNA PG4s most of the neighbouring genomic context is constrained by a complementary strand (Millevoi *et al.*, 2012). Mfe values obtained with the RNAfold software, which were used as an indication of the relative competitiveness of the neighbouring genomic context, were not efficient predictors of G4 folding for long transcripts. Moreover, it is important to note that this software cannot predict G4 folding. Recently, Lorenz and coworkers published new RNA folding algorithms taking G4s into consideration (Lorenz *et al.*, 2012). Thus we compared this new RNA folding algorithm's predictions to the cG/cC score predictions, for all of the candidates probed *in vitro* in this study. Most folding predictions were in agreement with our *in vitro* probing results. However, for the 14 candidates used in the cG/cC score challenge (**Table 4**), four that did not fold into G4s *in vitro* were nevertheless predicted to do so by the new RNA folding algorithm, representing as many false positives. The cG/cC score analysis of the first set of 12 long-transcript candidates permitted to evaluate the predictive

ability of different threshold values in terms of sensitivity and specificity (**Supplementary Table S6 in Annexe 2**). A threshold of 2.05 had 100% sensitivity and 83.3% specificity. Candidates with a cG/cC score > 2.05 are predicted to fold into G4s, whereas candidates with smaller scores are predicted to adopt canonical Watson-Crick structures. Using this cG/cC threshold on the second set of 14 PG4 candidates listed in **Table 4** (cG/cC scores were not used to establish the threshold), only three false positives were obtained that is a few less than with the new RNA folding algorithm. To increase specificity of the cG/cC score, a higher threshold value of 3.05 was selected. This new threshold yielded a total of only two discrepancies (i.e. one false positive and one false negative). Taken together, these results show that the new RNA folding algorithm is a fairly effective predictor of G4 folding, but that the cG/cC score can improve and refine predictions.

Future work should focus on assessing G4 folding of a larger set of RNA PG4s within their neighbouring genomic context in order to further refine the predictive power of cG/cC thresholds. Here, we considered RNA PG4 candidates with genomic context sizes of 15 nt and 50 nt. Investigation of larger genomic context sizes should help optimize predictive value. Even though not tested *per se* within the scope of this study, a determinable size limit must apply. Indeed, considering too much genomic context will only dilute the informative value of relative C and G contents. C and G contents nevertheless impact predictive power more strongly than does the exact context size. Therefore, considering a genomic context of 50 nt seems adequate, and fine-tuning of the context size still worthy of future investigations. Because of the intrinsic environmental differences between RNA and DNA PG4s, it is likely that their respective cG/cC thresholds are expected to differ. Window length, or the distance of genomic context considered as impacting G4 folding, is also expected to differ – that is to be smaller for DNA PG4s because of its double-strandedness. The cG/cC threshold of a given PG4 may also vary owing to its relative position within the genome. For example, PG4s located in coding versus non-coding regions. It is also not impossible that PG4s located in an otherwise unfavorable genomic context or position still form G4s *in vivo* when folding co-factors such as either proteins or small RNAs, bind to neighbouring inhibitory C-tracts as previously proposed [(Beaudoin et Perreault, 2010); S. Rouleau, JD Beaudoin and JP Perreault, unpublished data]. RNA G4 folding is decidedly more complex than meets the eye. For instance, the same transcript could perhaps generate alternative folding and non-folding

isoforms, in differing proportions depending on prevailing cellular conditions. Still another RNA PG4 candidate predicted as non-folding could still yield a small proportion of folding transcripts exerting biological effects despite predictions to the contrary. While the current state of knowledge makes such conjectures premature, future investigations will no doubt continue to shed fascinating insights into the nuances of G4 folding and the prediction thereof. With increasing evidence of 'atypical' G4 folding, turning to a predictive scoring system based on guanine density, instead of a standard algorithm, is gaining support and appears to be a suitable avenue to increase the accuracy of G4 folding predictions [(Eddy et Maizels, 2006 ; Huppert, 2008b ; Mukundan et Phan, 2013 ; Rawal *et al.*, 2006); J.L.Mergny, unpublished data]. Prediction of RNA G4 folding based on energy-based models such as the algorithm proposed by Lorenz and coworkers is still in its infancy (Lorenz *et al.*, 2012). Currently, relatively little knowledge exists about the energies driving G4 folding compared to canonical Watson-Crick structures. The cG/cC score developed here provides a convenient complementary tool to currently available G4 prediction softwares. Thus the cG/cC score can help discard incorrect folding predictions while essential and more accurate energy-based models are refined.

**Concluding remarks**

In summary, the PG4 genomic context, and especially consecutive cytosine content, appears to be one of the main criteria governing G4 folding of RNA molecules. The cG/cC score presented here is representative of this context and based on relative consecutive G to C contents. This helpful new parameter predicts RNA G4 folding with relatively high sensitivity and specificity. It is most useful when used in combination with current G4 prediction tools. The accurate prediction of RNA G4 folding is an essential step toward a better understanding of both their functional roles and biological importance. All of this is part of a much broader challenge aiming to uncover and comprehend the intricate regulatory mechanisms of RNA G4 folding underlying the biological processes of disease and health.

## SUPPLEMENTARY DATA

### Supplementary Datasets

Available at NAR online, URL : https://doi.org/10.1093/nar/gkt904

**Supplementary Dataset 1:** 5'-UTR database with cG/cC scores (AddSuppFiles-1 - xlsx file)
**Supplementary Dataset 1:** 3'-UTR database with cG/cC scores (AddSuppFiles-2 - xlsx file)

**Annexe 2**
**Supplementary Tables S1-S6**
**Supplementary Figures and Legends S1-S27**

## ACKNOWLEGMENTS

The authors thank Dominique Lévesque for technical assistance in the initial steps of this study and Monique Sullivan for edition of the manuscript. R.J. was the recipient of the CIHR Frederick Banting and Charles Best Canada Graduate Scholarship Master's Award. J.D.B. was the recipient of the CIHR Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Award. J.P.P. holds the Chaire de recherche de l'Université de Sherbrooke en Structure et Génomique de l'ARN and is a member of the Centre de Recherche Clinique Étienne-Le Bel.

## FUNDINGS

Canadian Institute of Health Research (CIHR) [MOP-44022 to J.P.P.]; Université de Sherbrooke [to the RNA group]. Funding for open access charge: Canadian Institute of Health Research (CIHR) [MOP-44022].

# ARTICLE 3 – THE FOLDING OF 5'UTR HUMAN G-QUADRUPLEXES POSSESSING A LONG CENTRAL LOOP

**Auteurs de l'article** : Jodoin, Rachel*, Bauer, Lubos*, Garant Jean-Michel*, Laaref, Abdelhamid Mahdi, Phaneuf Francis et Perreault, Jean-Pierre
*Contributions égales

**Statut de l'article** : Publié dans RNA (2014), vol. 20, no.7, p.1129-1141

**Avant-propos** : Rachel Jodoin a effectué les essais de cartographie *in vitro* de 5 des 8 candidats ainsi que les essais *in cellulo*. Lubos Bauer a analysé la banque de séquences potentielles de PG4 à longue boucle. Jean-Michel Garant, assisté par Abdelhamid Mahdi Laaref et Francis Phaneuf a effectué la cartographie *in vitro* des 3 autres candidats. L'article a été rédigé par Rachel Jodoin, Lubos Bauer, Jean-Michel Garant et Jean-Pierre Perreault.

**Résumé**

Les G-quadruplexes sont des structures à 4 brins répandues, qui sont adoptées autant par des régions G-riches d'ADN ou d'ARN, et qui sont impliquées dans des processus biologiques essentiels tels que la traduction des ARNm. Elles sont formées par l'empilement de 2 tétrades de G ou plus, liées entre elles par trois boucles. Malgré que la taille des boucles soit usuellement limitée à 7 nt dans la plupart des logiciels de prédiction de G-quadruplex, il a déjà été démontré que des séquences artificielles de G-quadruplex d'ADN contenant 2 boucles distales limitées à 1 nt chacune, avec une boucle centrale de 30 nt, pouvaient se replier *in vitro*. Ce rapport démontre que de telles structures avec une longue boucle centrale sont actuellement retrouvées dans les 5'UTR des ARNm humains. Premièrement, 1453 G-quadruplex potentiels (PG4) ont été identifiés à l'aide d'une recherche bio-informatique de séquences correspondant à 2 boucles distales de 1 nt et une boucle centrale d'une longueur de 2 à 90 nt. Deuxièmement, des expériences de cartographies *in-line in vitro* ont confirmé et caractérisé le repliement de 8 candidats possédant des boucles centrales longues de 10 à 70 nt. Finalement, l'effet biologique sur les niveaux d'expression d'ARNm de quelques G-

quadruplexes possédant une longue boucle centrale a été étudié *in cellulo* par l'utilisation d'un gène rapporteur luciférase. Clairement, la définition actuelle d'une séquence formant un G-quadruplex est trop conservatrice et doit être étendue afin d'inclure de plus longues boucles centrales. Ceci augmente grandement le nombre de PG4 attendu dans le transcriptome. La considération de ces nouveaux candidats pourrait aider à l'élucidation des potentielles implications biologiques importantes de la structure G-quadruplex.

**Abstract**

G-quadruplexes are widespread four-stranded structures that are adopted by G-rich regions of both DNA and RNA and are involved in essential biological processes such as mRNA translation. They are formed by the stacking of two or more G-quartets that are linked together by three loops. Although the maximal loop length is usually fixed to 7 nt in most G-quadruplex predicting software, it has already been demonstrated that artificial DNA G-quadruplexes containing two distal loops that are limited to 1 nt each and a central loop up to 30 nt long are likely to form *in vitro*. This report demonstrates that such structures possessing a long central loop are actually found in the 5'-UTRs of human mRNAs. Firstly, 1453 potential G-quadruplex forming sequences (PG4s) were identified through a bioinformatic survey that searched for sequences respecting the requirement for two one nt long distal loops and a long central loop of 2–90 nt in length. Secondly, *in vitro* in-line probing experiments confirmed and characterized the folding of eight candidates possessing central loops of 10–70 nt long. Finally, the biological effect of several G-quadruplexes with a long central loop on mRNA expression was studied *in cellulo* using a luciferase gene reporter assay. Clearly, the actual definition of G-quadruplex forming sequences is too conservative and must be expanded to include the long central loop. This greatly expands the number of expected PG4s in the transcriptome. Consideration of these new candidates might aid in elucidating the potentially important biological implications of the G-quadruplex structure.

## INTRODUCTION

Guanine-rich nucleic acid sequences can fold into a well-known tetrahelical structure called G-quadruplex. The basic building blocks of the G-quadruplex core are two or more G-quartets, which are planar arrangements of four guanines held together by Hoogsteen hydrogen bonds pairing (Gellert *et al.*, 1962). The structure is formed by the stacking of the G-quartets on top of each other and is further stabilized by the binding of monovalent ions, especially $Na^+$ and $K^+$. A typical intramolecular G-quadruplex forming sequence is composed of four tracts of two or more consecutive guanines (G-tracts) which are interspersed by three loops of variable lengths and nucleotide compositions. The stability of the structure is affected by several features including: the number of G-quartets, the possibility of bulge formation, the type and concentration of monovalent cations in solution, the sequence of the nucleic acid molecule itself and the length of the loops composing the G-quadruplex (Burge *et al.*, 2006; Mukundan et Phan, 2013). Several studies focused on the bioinformatic analysis of G-quadruplexes in the human genome confirmed the presence of a significant number of potential G-quadruplex forming sequences (PG4s) in various biologically relevant regulatory regions such as the promoter elements of genes, telomeres and the UTRs of mRNAs (Beaudoin et Perreault, 2010, 2013 ; Eddy et Maizels, 2006 ; Huppert *et al.*, 2008 ; Huppert et Balasubramanian, 2005). The existence of RNA G-quadruplexes in human cells was recently confirmed using a structure specific antibody (Biffi *et al.*, 2014a). A significant number of studies have linked G-quadruplexes to important biological processes, including: mRNA splicing, polyadenylation, translation repression and localization (Beaudoin et Perreault, 2010, 2013 ; Bugaut et Balasubramanian, 2012 ; Marcel *et al.*, 2011 ; Millevoi *et al.*, 2012 ; Shafer et Smirnov, 2000), thus rendering them interesting potential therapeutic targets (Collie et Parkinson, 2011 ; Marcel *et al.*, 2011 ; McLuckie *et al.*, 2013 ; Patel *et al.*, 2007).

Biophysical studies have confirmed that RNA G-quadruplexes are generally thermodynamically more stable than their DNA counterparts (Halder et Hartig, 2011 ; Joachimi *et al.*, 2009 ; Saccà *et al.*, 2005 ; Zhang *et al.*, 2011a). Moreover, RNA G-quadruplexes are restricted to adopting a parallel configuration caused by the stronger preference for an *anti*-conformation of the glycosidic bond between the ribose and guanine moieties (Halder et Hartig, 2011). Considerable effort has been spent trying to understand

the principles which govern the folding of G-quadruplexes (Hardin *et al.*, 2000; Karsisiotis *et al.*, 2013; Xue *et al.*, 2011). Numerous articles have explored the contributions of the composition and length of the loops on the formation and topology of both DNA and RNA G-quadruplexes (Guédin *et al.*, 2009, 2010 ; Hazel *et al.*, 2004 ; Koirala *et al.*, 2013 ; Kwok *et al.*, 2013 ; Olsen *et al.*, 2009 ; Pandey *et al.*, 2013 ; Rachwal *et al.*, 2007b ; Risitano et Fox, 2004 ; Zhang *et al.*, 2011a), and some general conclusions regarding the loops have emerged. Firstly, in contrast to DNA, the topology of RNA G-quadruplexes is always parallel and independent of the loop length and sequence (Pandey *et al.*, 2013; Zhang *et al.*, 2011a). Secondly, the stability of both DNA and RNA quadruplexes and the length of the loops are inversely related. In other words, G-quadruplexes with shorter loops exhibit higher stability than those with longer loops. However, it is very important to note that this holds true only for sequences with shorter loops. If a G-quadruplex structure harbours longer loops (>20 nt) a plateau is attained and the stability becomes less dependent on the loop length (Guédin *et al.*, 2010; Pandey *et al.*, 2013). Moreover, it was established that if a very long central loop is accompanied by two short loops comprised of a single nucleotide each, the stability of the G-quadruplex was still relatively high, exceeding the physiological temperature (Guédin *et al.*, 2010; Pandey *et al.*, 2013). The majority of these studies were conducted on artificial DNA sequences in which the length of the loops did not exceed 30 nt. Despite the numerous studies, the issue of longer loops occurring in natural RNA G-quadruplexes still remains poorly explored. In accordance with these conclusions, it seemed plausible that 5'-UTR RNA G-quadruplexes with longer loop lengths could be stable enough to be formed and retrieved in the human transcriptome. If this is indeed the case, they could act as translational repressors (Beaudoin et Perreault, 2010 ; Bugaut et Balasubramanian, 2012). To verify these assumptions, a database of 1453 human 5'-UTR PG4s composed of two distal loops of length of 1 nt and a central loop of varying lengths, ranging from 2 up to 90 nt was constructed. The folding of eight representative PG4s with different central loop lengths was confirmed *in vitro*, and in some cases *in cellulo*. All of the PG4s investigated defy the classical algorithm respecting 7-nt long loops only.

## RESULTS

### Database of G-quadruplexes possessing a long central loop

Initially, PG4s were searched using the algorithm Gx-$N_1$-Gx-$N_{2\text{-}90}$-Gx-$N_1$-Gx, were G stands for guanine, N for any nucleotide (A, U, C and G) and $x \geq 3$. This *in silico* analysis of the human 5'-UTRs yielded 1453 PG4 sequences with central loops ranging from 2 to 90 nt accompanied by 1-nt-long distal loops. Out of the 1453 PG4s, 1232 were comprised of central loops >8 nt, therefore deviating from the widely used search algorithm. The analysis of the constructed database permitted the observation of some interesting tendencies of the PG4 sequences found in human 5'-UTRs. Comparing the lengths of the central loop revealed that PG4s with shorter loops were more frequent (**Figure 26A**) and that there was a tendency showing that the longer the loop, the fewer the number of PG4s retrieved. The positions of the PG4 within the 5'-UTR demonstrated that they tend to localize at the 5'-extremity of the 5'-UTR (**Figure 26B**), which is in agreement with the work on RNA G-quadruplex corresponding to the canonical definition (Huppert *et al.*, 2008).

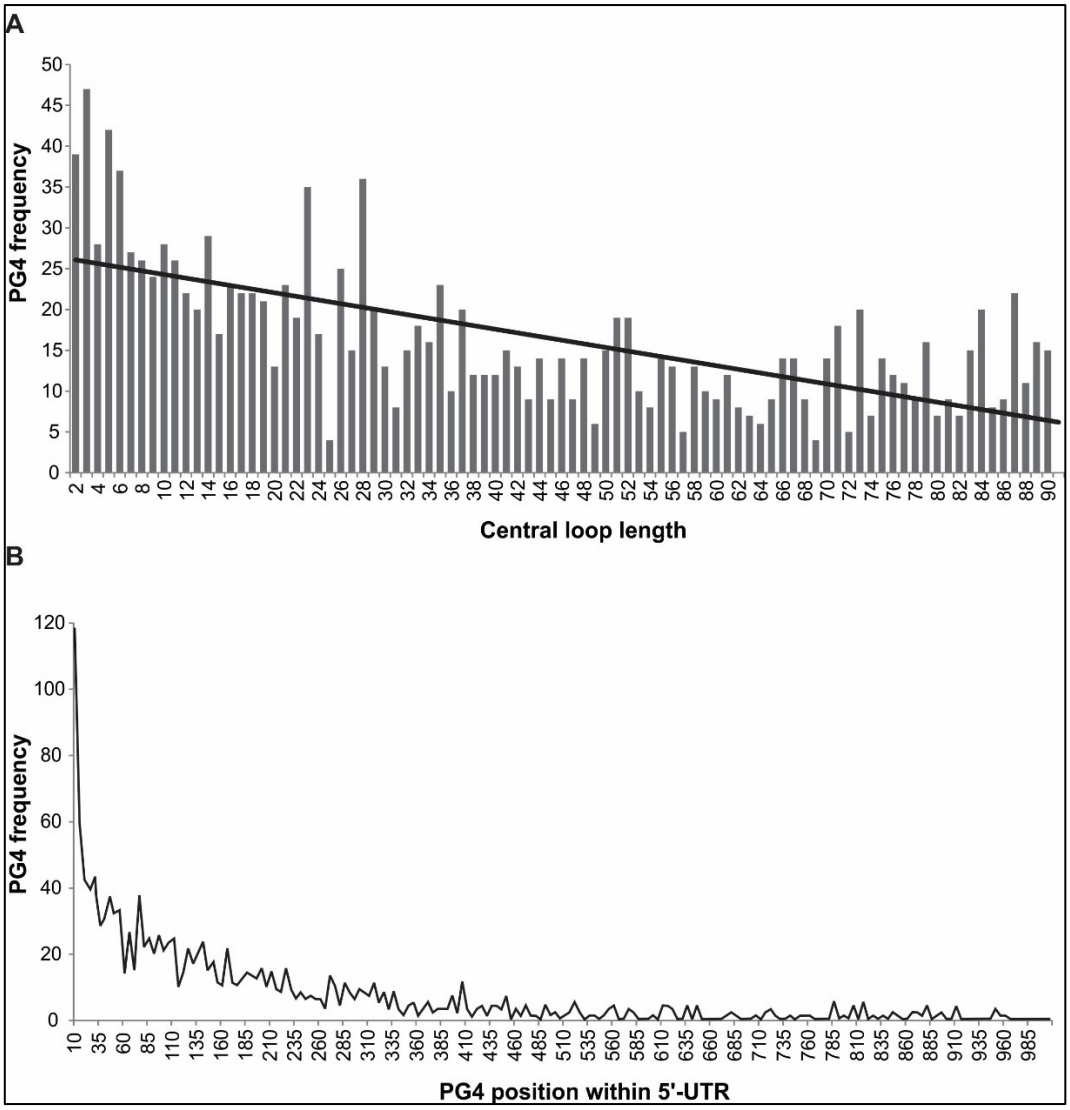



**Figure 26** – Distribution of the central loop lengths of the PG4 and long loop PG position within 5'UTR

(A) Incidence of potential G-quadruplexes (PG4s) possessing central loops of varying lengths in a human 5'-UTR database. (B) Position of PG4s within the 5'-UTR.

### In vitro folding of potential G-quadruplex-forming sequences possessing a long central loop

Representative, natural 5'-UTR PG4 sequences with variable central loop lengths were chosen from the database (**Table 5**) and subjected to in-line probing experiments to verify their ability to fold into G-quadruplex structures *in vitro.* This technique has been very successfully used to follow the formation of G-quadruplexes located in both 5'- and 3'-UTRs of RNA transcripts (Beaudoin et Perreault, 2010, 2013). In addition, a step-by-step

methodology of the whole in-line probing protocol, including the design of the PG4s, performing of the experiments, and the evaluation of the data has already been described in detail (Beaudoin *et al.*, 2013). Briefly, this assay makes use of the natural instability of RNA to elucidate secondary structure characteristics. For instance, when a PG4 sequence adopts an intramolecular G-quadruplex structure, the nucleotides in the loops should bulge out of the RNA's structure and should therefore be more susceptible to spontaneous non-enzymatic cleavage of their phosphodiester bonds, a process that is favored by the presence of magnesium ions. To render the analysis more biologically relevant, extra 15 nt sequences were added to both ends of the PG4 sequence. This permitted observation of the formation of the G-quadruplex structure in its broader genomic context. In addition to the wild-type (wt) PG4 version, a mutated version in which some key guanines were substituted for by adenines (G/A-mut) was synthesized in each case. The G/A-mutant served as a negative control for G-quadruplex formation as it possessed only minor changes in its RNA sequence compared to that of the wt. Knowing that $Li^+$ cations are unable to stabilize the G-quadruplex structure, due to their small size, another layer of control was added and the in-line reactions were performed in the presence of 100 mM of both $K^+$ and $Li^+$ to favor and disfavor, respectively, the formation of G-quadruplexes.

**Table 5** Characteristics of selected PG4 candidates

| 5'-UTR | | | Potential RNA G-quadruplex (PG4) | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Gene** | **Refseq name** | **Length** | **Position within UTR** | **Length** | **G-tract** | **L1*** | **L2* (length)** | **L3*** |
| BAG1 | NM_004323 | 89 | 9 | 28 | GGG | C | GCUGGGAAGUAGUC (14) | C |
| HIRA | NM_003325 | 220 | 102 | 25 | GGG | C | CGGCGGCCGGA (11) | C |
| CTGLF6 | XM_001716810 | 1750 | 1274 | 70 | GGGG | A | UGGCAGGCAGGGUGGGGC ACUGUGAGGUGUCGGGGA GGGCAUUGUGAAGUGU (52) | U |
| TOM1L2 | NM_001033551 | 157 | 15 | 50 | GGGG | C | CCAAAGGCCCUAAGCUCG GCGUUCCAGAGAGU(32) | A |
| CBX1 | NM_006807 | 481 | 137 | 47 | GGG | C | GCGCGAAUCCUGAGCCAG AGACUGAGUGCUUGG (33) | U |
| APC | NM_001127511 | 380 | 29 | 44 | GGG | C | GUGUGGCCGCCGGAAGCC UAGCCGCUGCUC (30) | G |
| MDS1 | NM_004991 | 396 | 155 | 85 | GGG | A | AGAGAGAGUGAAAGAAGA AAAUACAGAGAGUGAGUG UGUGGAAGAGAGAGAGAA ACAGGAGAGAAACAGGA (71) | A |
| LRRC37A3 | NM_199340 | 531 | 407 | 83 | GGG | C | AUUGUGACAUAAGAGUGC CCUGGUGACAUGGAGCAG AUCUGUGGCAUAAAUAAA GGUGUCAUAAAGACA (69) | C |

* L1, L2 and L3 represent the first, second and third loop of the G-quadruplex respectively

*BAG-1*

Initially the PG4 found in the 5'-UTR of the human BAG1 mRNA was chosen from the database to assess its ability to form a G-quadruplex possessing a long central loop. The BAG1 PG4 was predicted to be comprised of 28 nt with a central loop of 14 nt and forming a G-quadruplex with three G-quartet layers (**Table 5**). The analyzed sequence of BAG1 is shown in **Figure 27A**. The boxed nucleotides represent the PG4, and the tracts of guanines predicted to be involved in the formation of the G-quadruplex are underlined. A typical autoradiogram for an in-line probing analysis of both the wt and G/A-mut versions of the BAG1 PG4 is illustrated in **Figure 27B**. Differences in the intensities of some bands were observed at several positions of the wt PG4 in the presence of 100 mM KCl as compared in the presence of 100 mM LiCl. More specifically, the bands corresponding to the nucleotides found in the predicted loops that are located between the guanine tracts (i.e. C14, G18, C19, U20, A24 and C35) became more intense only for the wt version in the presence of KCl. In

addition, the inability of the G/A-mutants to fold into a G-quadruplex structure was confirmed, regardless of the type of the cation used. To quantitatively evaluate the in-line probing analysis, the intensity of each band in the $K^+$ lane was divided by that of the corresponding band in the $Li^+$ lane. The retrieved $K^+/Li^+$ ratios for each band were further used to create bar graphs (**Figure 27C**) with the nucleotide sequence plotted on the *y*-axis and the intensity ratios on the *x*-axis. A nucleotide was considered significantly more accessible when this ratio was higher than an arbitrarily fixed threshold of 2. As expected, the ratios of the nucleotides located between the tracts of guanines were superior to the arbitrary threshold, suggesting that BAG1 forms a RNA G-quadruplex with a 14-nt-long central loop. Since the sequence of the central loop contains an additional G-tract, it is reasonable to assume that it might be involved in the formation of alternative G-quadruplexes. The extra G-tract could provide multiple folding scenarios and support the formation of various G-quadruplex structures. In this case the resulting cleavage pattern would reflect the sum of multiple G-quadruplex species present in solution during the 40-h-long incubation procedure. To get insight into this hypothesis, a mutant BAG1 was constructed. Guanines G22 and G23, which are located in the central loop, were changed to adenines. The in-line probing was performed on this mutated sequence, followed by a quantitative analysis of the bands. The significant increase in the intensity of nucleotide C14 located between the first and second G-tract implies that a new equilibrium was established and that only one species with a long central loop was favored (**Figure 27D**).

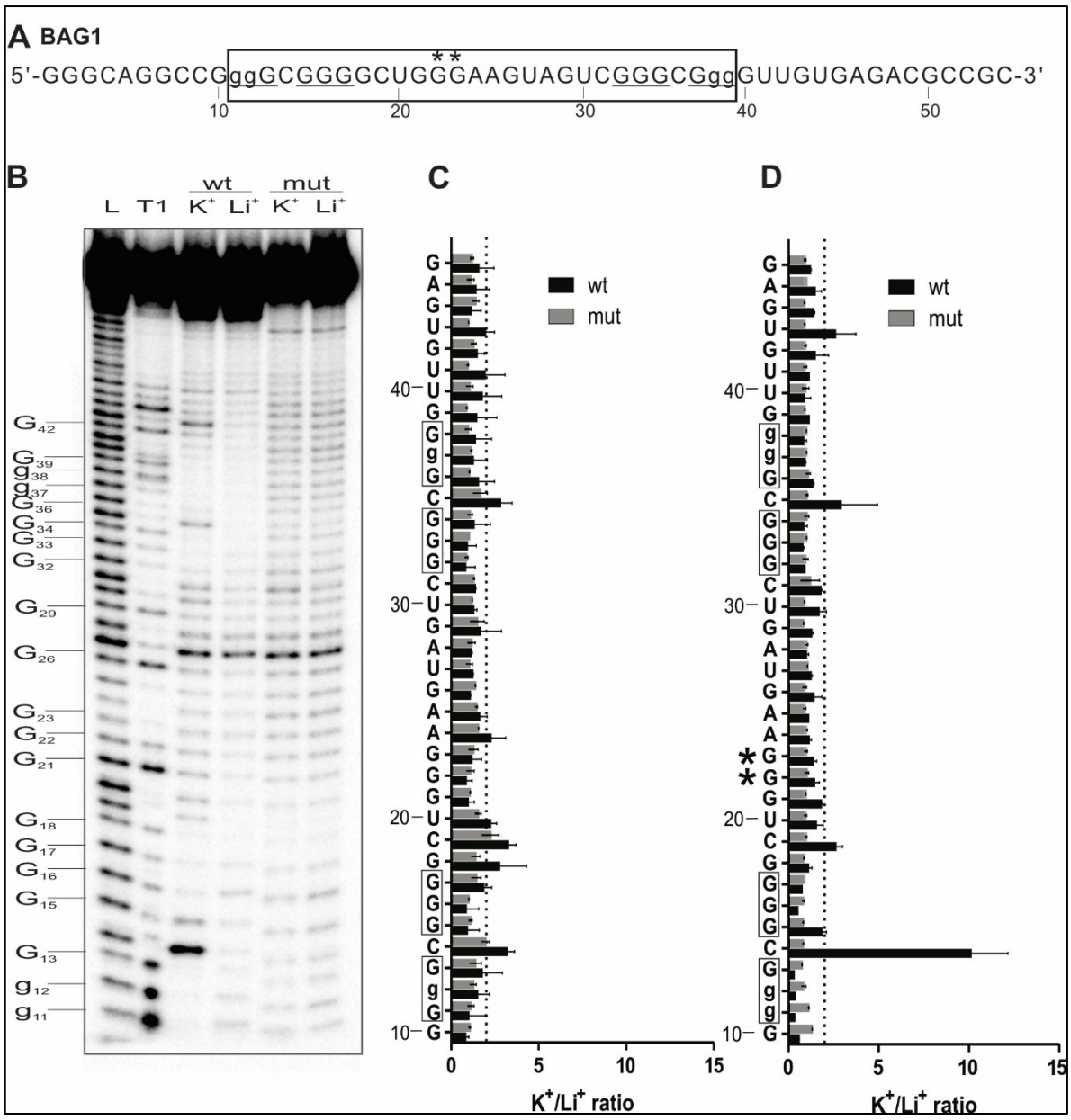



**Figure 27** – In-line probing results of the BAG1 PG4 candidate which possesses a 14-nt-long central loop.

(A) Nucleotide sequence of the characterized BAG1 wt transcript. The lowercase guanines (g) correspond to those substituted for by adenines in the G/A-mutant versions. Guanines mutated in the central loop are denoted by asterisks (*). Underlined G-tracts indicate the nucleotides predicted to be involved in the G-quadruplex formation. The boxed sequence denotes the predicted PG4. (B) Autoradiogram of a 10% denaturing (8 M urea) polyacrylamide gel of the in-line probing of both the 5'-labelled BAG1 wt and G/A-mutant PG4 versions performed in the presence of 100 mM of either LiCl or KCl. The L and T1 lanes indicate the alkaline hydrolysis and ribonuclease T1 mapping lanes, respectively. The positions of the guanines are indicated on the left of the gel. The lowercase guanines were converted to adenines in the mutant version. (C,D) $K^+/Li^+$ ratios of the band intensities of the BAG1 wt and the G/A-mutant for each nucleotide. The $K^+/Li^+$ ratios are shown in dark grey for BAG1

wt and in light grey for the BAG1 G/A-mutant. The boxed guanines represent the predicted G-tracts. The dotted line represents the 2-fold threshold that denotes a significant gain in flexibility. The sequence is indicated on the *y*-axis. The lower case G's shown on the *y*-axis are mutated to A's in the mutant version. The asterisk (*) indicates guanines mutated to adenines in the central loop. Each bar represents the average of two independent experiments, and the error bars represent the standard deviations.

*HIRA*

Based on observations made on BAG1 and the guanine tract located in the central loop, further investigation focused on HIRA (**Table 5**), a candidate predicted to fold into a G-quadruplex structure composed of three G-quartets with an 11-nt-long central loop harboring three guanine doublets (identified with asterisks in **Figure 28A**). In the presence of KCl the wt sequence displayed an in-line cleavage pattern typical of the formation of multiple G-quadruplex species. Besides the nucleotides which were predicted to be the first and third single nucleotide loops of the PG4 (C28, C46), additional accessible sites superior to the arbitrary threshold were identified in the long central loop (C32, G33, C35, G36, C38, and C39). The localization of residues between doublets of guanines (C32, C35, C38, and C39) strongly supports the existence of an alternative G-quadruplex consisting of two G-quartets, the minimum requirement for the structure. Because of the presence of multiple G-tracts, and of several folding combinations, it is complicated to evaluate which of them are involved in the formation of particular G-quadruplexes. Moreover, a guanine doublet located just after the predicted PG4 might be also involved in the formation of an alternative G-quadruplex as evidenced by the superior accessibility of nucleotide U50. To prove the existence of an alternative G-quadruplex topology composed of two G-quartets, G to A mutations were introduced into the three guanine doublets. The cleavage susceptibility of the nucleotides located between the doublets did indeed decrease under the threshold, and the folding of the originally predicted PG4 with a long central loop was promoted, as is indicated by the increased cleavage ratios of both C28 and C46 (**Figure 28C**). The mutation of only the first guanine doublet reduced the in-line cleavage susceptibility of nucleotides C32 and C35 (**Figure 28D**). This result is in concordance with the previous mutation of all guanine doublets, and reinforces the hypothesis that the doublets offer alternative folding pathways.

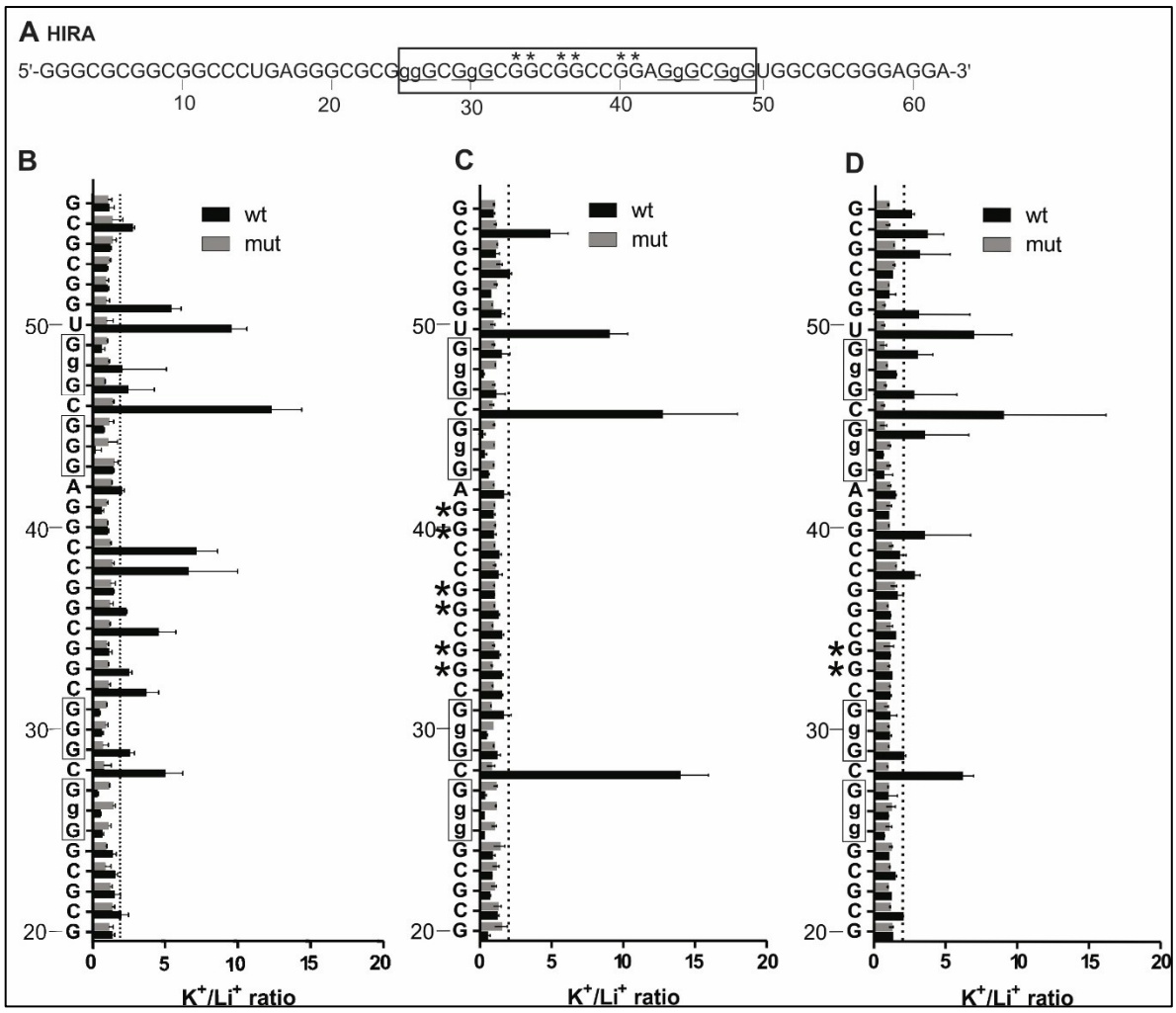



**Figure 28** – In-line probing results of the HIRA PG4 candidate which possesses an 11-nt central loop.

(A) Nucleotides sequence of the characterized HIRA wt transcript. The lowercase guanines (g) correspond to those substituted for by adenines in the G/A-mutant version. Guanines mutated in the central loop are denoted by asterisks (*). Underlined G-tracts indicate the nucleotides predicted to be involved in the G-quadruplex formation. The boxed sequence denotes the predicted PG4. (B-D) $K^+/Li^+$ ratios of the band intensities of the HIRA wt and the G/A-mutant *in vitro* G-quadruplex version for each nucleotide. The $K^+/Li^+$ ratios are shown in dark grey for the HIRA wt and in light grey for the HIRA G/A-mutant. The boxed guanines represent the predicted G-tracts. The lower case G's shown on the *y*-axis are mutated to A's in the mutant version. The dotted line represents the 2-fold threshold that denotes a significant gain in flexibility. The nucleotide sequence is indicated on the *y*-axis. The asterisk (*) indicates guanines mutated to adenines in the central loop. Each bar represents the average of two independent experiments, and the error bars represent the standard deviations.

*CTGLF6*

The next candidate to be examined was CTGLF6, a sequence capable of folding into multiple G-quadruplex species with different loop lengths depending on which combination of G-tracts is considered to be involved in the formation of a particular structure (**Table 5**). The predicted PG4s harboring 10-, 16- and 14-nt-long central loops are denoted in solid, dotted and dashed boxes, respectively (**Figure 29A**). In addition, the PG4 consisting of the first two G-tracts of the first PG4 (solid box) and the last two G-tracts of the third PG4 (dashed box) with a long 52-nt central loop must also be considered. Furthermore, the existence of G-quadruplex subunits arranged in tandem between the first (solid) and third (dashed) PG4 intercalated with the central loop of the second PG4 (dotted) also seems to be a plausible possibility. Several mutations were performed to modulate the folding towards one specific structure by impairing the participation of specific G-tracts in the formation of G-quartets. The in-line probing pattern of the wt sequence in the presence of $K^+$ corresponds to the formation of two consecutively arranged G-quadruplexes, as expected (PG4 in the solid and dashed boxes, respectively). Although it is important to note that the accessibility of the nucleotides located between the G-tracts of the G-quadruplex located at the 3'-end is on the edge of the arbitrarily defined threshold. The first series of mutations was introduced with the intention of impairing the formation of the G-quadruplex located in the 5'-end. A decrease in the cleavage ratios for nucleotides A17 and U34 in the single stranded loops was observed. On the other hand, the intensity ratios of residues A57 and U78, situated amid the G-tracts of the 3'-end G-quadruplex, was slightly increased, indicating the promotion of this structure (**Figure 29B**). The same kind of behavior, but in an inverted order, was observed when the G-quadruplex situated at the 3'-end was mutated so as to abolish its formation. Specifically, the decreased cleavage ratios of nucleotides A57 and U78 in the first and third loops of the 3'-end G-quadruplex was accompanied by an increase in the cleavage ratios of the nucleotides located in the short loops (nucleotides A17 and U34) of the 5'-end G-quadruplex (**Figure 29C**). The last mutation performed was designed to promote the formation of a G-quadruplex structure in the center of the sequence (PG4 in the dotted box) by impairing the first two guanine tracts in the 5'-end PG4 (solid box) and the last two G-tracts in the 3'-end PG4 (dashed box). The resulting structure confirmed expectations, as both the first (U34) and

third loops (A57) exhibited higher cleavage ratios as compared to the wt sequence (**Figure 29D**).

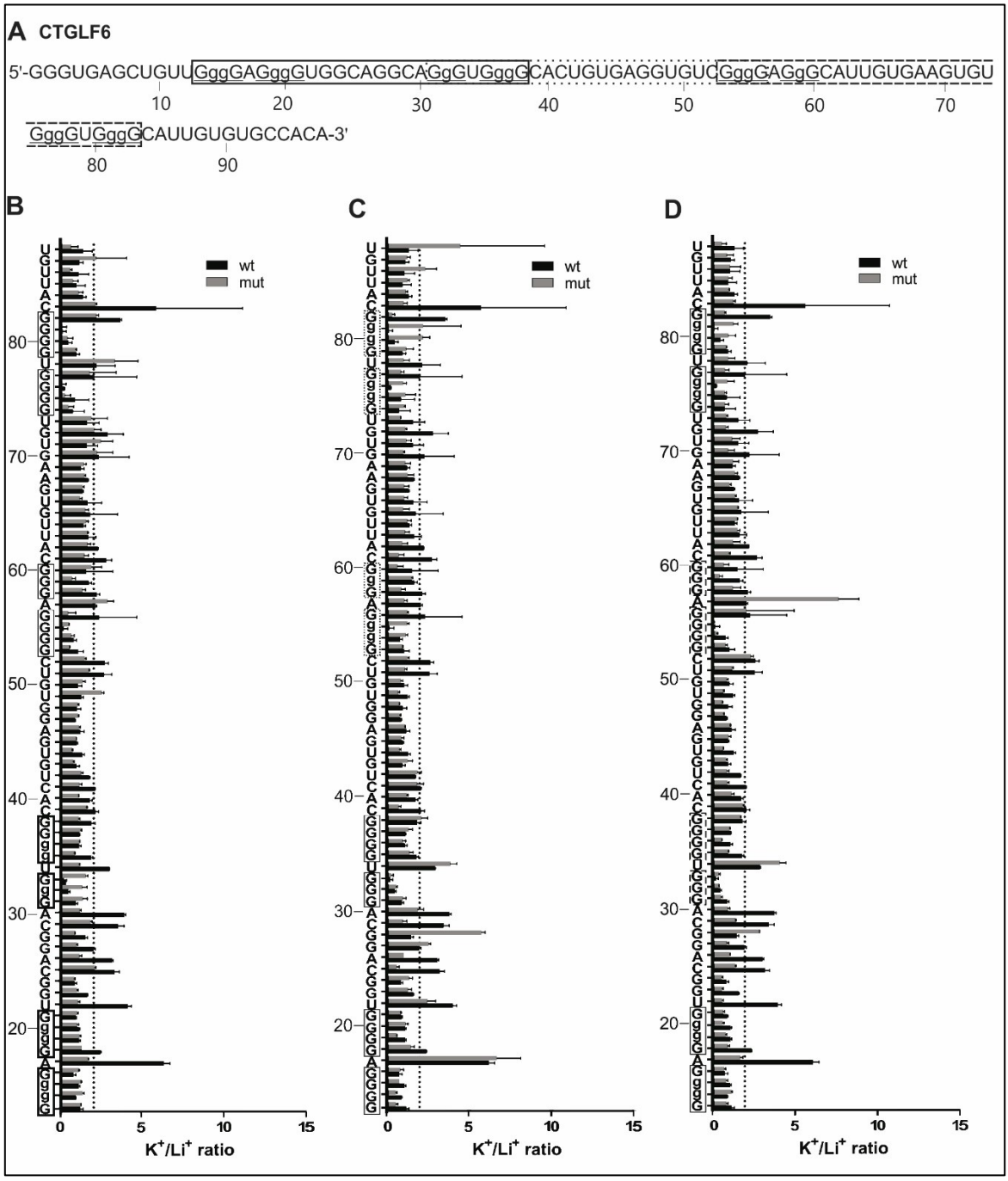



**Figure 29** – In-line probing results of the CTGLF6 PG4 candidate showing three overlapping PG4s, possessing a 10-, 16-, or 14-nt central loops.

(A) Nucleotide sequence of the characterized CTGLF6 wt transcript. The lowercase guanines (g) correspond to those substituted for by adenines in the G/A-mutant version. Underlined G-tracts

indicate the predicted nucleotides involved in the G-quadruplex formation. The boxed sequences in different frames denote the predicted PG4s. (B-D) $K^+/Li^+$ ratios of the band intensities of the CTGLF6 wt and the different G/A-mutants *in vitro* G-quadruplex versions for each nucleotide. (B) CTGLF6 wt and 5'-end G/A-mutant, (C) CTGLF6 wt and 3'-end G/A-mutant and (D) CTGLF6 wt and 5',3'-end G/A-mutant. The $K^+/Li^+$ ratios are shown in dark grey for the CTGLF6 wt and in light grey for the different CTGLF6 G/A-mutants. The boxed guanines represent the predicted G-tracts. The dotted line represents the 2-fold threshold that denotes a significant gain in flexibility. The nucleotide sequence is indicated on the *y*-axis. The lower case G's shown on the *y*-axis are mutated to A's in the mutant version. Each bar represents the average of two independent experiments, and the error bars represent the standard deviations.

*TOM1L2, CBX1, and APC*

Three candidates, namely TOM1L2, CBX1 and APC containing 32-, 33- and 30-nt-long central loops, respectively, were examined to confirm their ability to fold into RNA G-quadruplex structures (**Table 5**). With the exception of APC, the wt candidates displayed the typical banding patterns corresponding to the formation of G-quadruplexes in the presence of KCl. As expected, the superior cleavage ratios of nucleotides C22 and A63 in case of TOM1L2 and C22 and U62 in case of CBX1, which are located between the G-tracts, confirmed the initial observations (**Figure 30A, B**). The increased cleavage ratio at position G21 of TOM1L2 suggests that this nucleotide ends up in the loop with C22. This indicates that a G-quadruplex with a first loop of two nucleotides is formed (**Figure 30B**). In the case of APC, the inferior cleavage ratio of G56, which is located between the last two G-tracts, did not support the conclusion that a G-quadruplex is formed by this PG4 (**Figure 30D**). None of the three PG4s described contained extra G-tracts located in the central loop, nor in the 15 nt long overhangs flanking both sides of the predicted PG4s. This feature simplifies the evaluation and the interpretation of the data due to the absence of multiple G-quadruplex species.

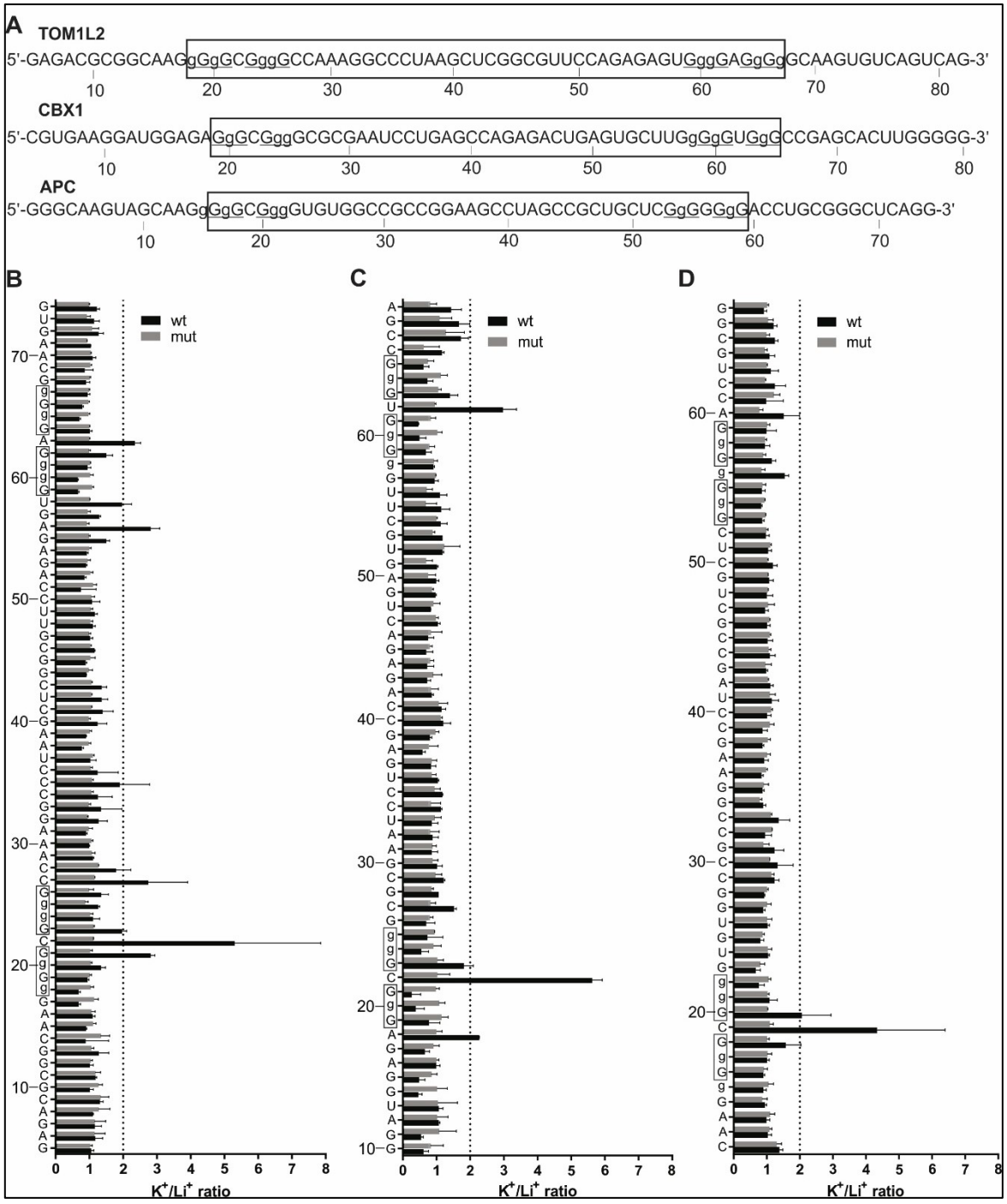



**Figure 30** – In-line probing results of the TOM1L2, CBX1, and APC, PG4s possessing centra loops of 32-, 33-, and 30-nt, respectively.

(A) Nucleotide sequences of the characterized wt transcripts. The lowercase guanines (g) correspond to those substituted for by adenines in the G/A-mutant version. Underlined G-tracts indicate the nucleotides predicted to be involved in the G-quadruplex formation. The boxed sequences denote the predicted PG4. (B-D) $K^+/Li^+$ ratios of band intensities of the wt and G/A-mutant for each nucleotide. (B) TOM1L2, (C) CBX1, and (D) APC. The $K^+/Li^+$ ratios are shown in dark grey for the wt and in

light grey for the G/A-mutant. The boxed guanines represent the predicted G-tracts. The dotted line represents the 2-fold threshold that denotes a significant gain in flexibility. The nucleotide sequence is indicated on the *y*-axis. The lower case G's shown on the *y*-axis are mutated to A's in the mutant version. Each bar represents the average of two independent experiments, and the error bars represent the standard deviations.

*MDS1 and LRRC37A3*

Lastly, two PG4s with central loops composed of 71- (MDS1) and 69- (LRRC37A3) nt (**Table 5**) were analyzed by in-line probing. In both cases, the wt sequences displayed exclusive higher $K^+/Li^+$ cleavage ratios of the residues predicted to be found in the single-stranded loops located between the G-tracts. More precisely, these residues correspond to nucleotides A24 and A103 for MDS1 and C17 and C93 for LRRC37A3. In comparison, the G/A mutants did not pose such characteristics, regardless of whether incubation was performed in the presence of LiCl or KCl (**Figure 31**). The assumption in the lack of cleavage difference for the nucleotides of the central loop could be explained by the presence of the same structure in both G-quadruplex favorable and unfavorable conditions. This argument is further supported by SHAPE probing experiments of the nucleotides located in the long central loop (data not shown). The SHAPE banding patterns obtained for the wt and G/A-mutant constructs were identical, indicating highly similar if not identical structures. The exceptionally long central loop of these PG4s sets a new limit of what might be still considered as an *in vitro* G-quadruplex forming sequence.

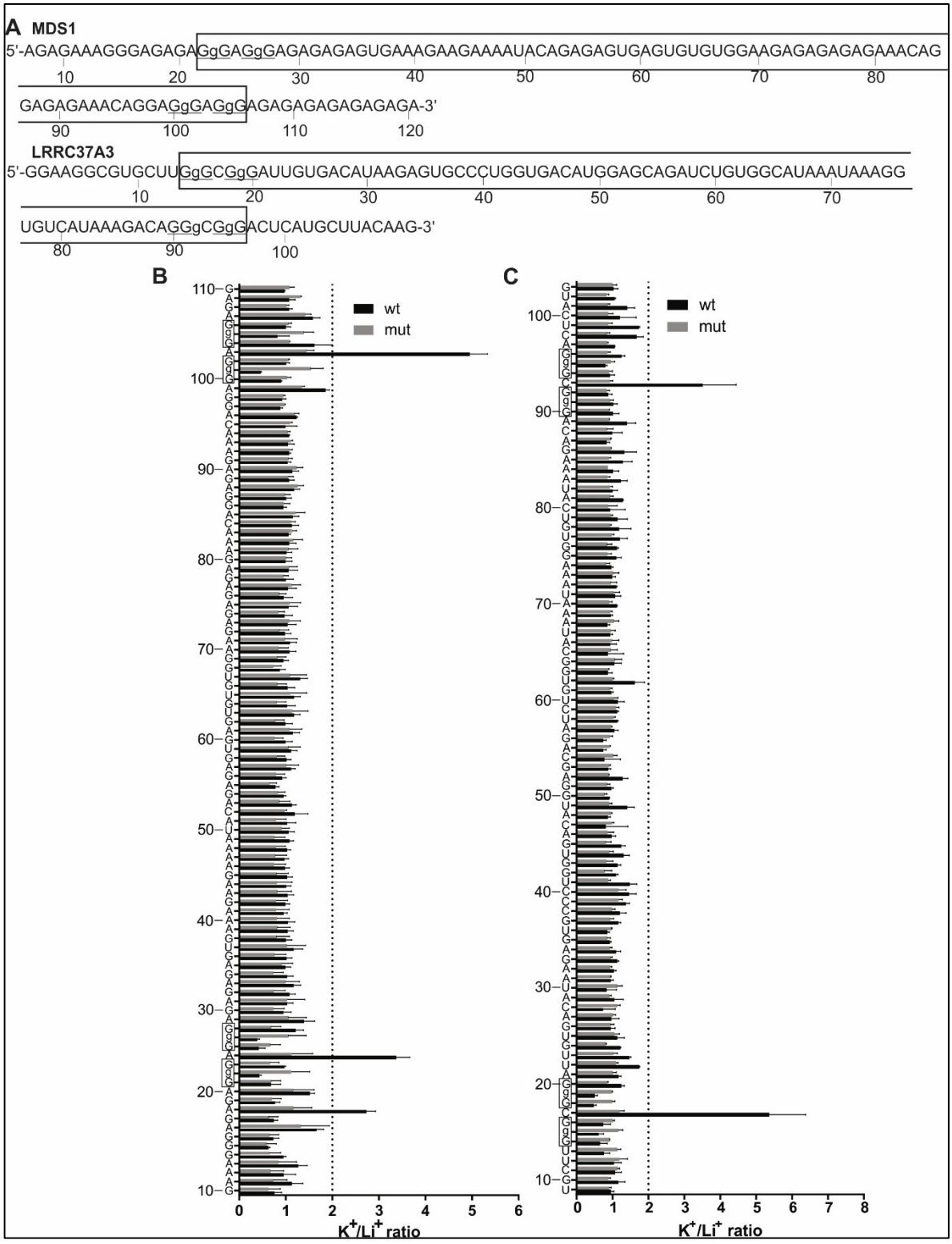



**Figure 31** – In-line probing results of the MDS1 and LRRC27A3 possessing central loops of 71- and 69-nt, respectively.

(*Legend Figure 31*) In-line probing results of the MDS1 and LRRC37A3 possessing central loops of 71- and 69-nt, respectively. (A) Nucleotide sequences of the characterized wt transcripts. The lowercase guanines (g) correspond to those substituted for by adenines in the G/A-mutant version. The underlined G-tracts indicate the nucleotides predicted to be involved in the G-quadruplex formation. The boxed sequences denote the predicted PG4. (B, C) $K^+/Li^+$ ratios of band intensities of the wt and the G/A-mutant for each nucleotide. (B) MDS1, (C) LRRC37A3. The $K^+/Li^+$ ratios are shown in dark grey for the wt and in light grey for the G/A-mutant. The boxed guanines represent the predicted G-tracts. The dotted line represents the 2-fold threshold that denotes a significant gain in flexibility. The nucleotide sequence is indicated on the *y*-axis. The lower case G's shown on the *y*-axis are mutated to A's in the mutant version. Each bar represents the average of two independent experiments, and the error bars represent the standard deviations.

**In cellulo folding of G-quadruplexes possessing a long central loop**

Encouraged by the results indicating that these G-quadruplex structures were folded *in vitro*, the next step was to verify their biological relevance by investigating their folding *in cellulo*. Multiple RNA G-quadruplex motifs located in the 5'-UTR of genes are reported to inhibit translation (Millevoi *et al.*, 2012). With the use of dual-luciferase reporter assays, we investigated whether or not some of the above candidates possessing unusual long central loops of 11-, 30-, and 70-nt could trigger the same effect. The complete (full-length) 5'-UTRs of the candidates were inserted upstream of the Renilla luciferase (Rluc) reporter gene. The levels of Rluc expression, normalized over the control Firefly luciferase (Fluc) expression, were compared between the wt constructs and the different G/A-mutants. The mutations used were the same as for the *in vitro* in-line probing assays. **Figure 32A** presents a schema of the different constructs of the HIRA candidate. To facilitate the comparison between each construction, and between different candidates, the luciferase activity of each construct was normalized over its corresponding G-tracts G/A-mutant and reported as a percentage. As expected for G-quadruplex formation, luciferase activity of the HIRA wt construct was reduced almost 90% as compared to the G/A-mutant construct which cannot adopt a G-quadruplex (**Figure 32B**). A smaller, but still important decrease of approximately 80% was also observed for the construction with G/A-mutation in the central loop. Accordingly to the *in vitro* in-line probing results, the HIRA wt construct could adopt multiple G-quadruplexes depending on the different combinations of the G-tracts used to form the structure. It seems that this pool of variable G-quadruplexes with different loop lengths and G-tracts has a higher detrimental impact on the expression of the luciferase gene than does a pool where a G-quadruplex with a long central loop is dominant, as is the case for the central loop G/A-mutant construct. However, in both cases, G-quadruplexes were folded *in cellulo*. Similar

results were observed for both the APC and the TOM1L2 candidates which possess central loops of 30- and 32-nt, respectively. Decreases in the luciferase activities of ~40% for APC and of ~75% for TOM1L2, due to G-quadruplex formation were observed (**Figure 32C, D**). Data obtained from *in cellulo* experiments with APC are in disagreement with the *in vitro* results, which did not unambiguously confirm the formation of a G-quadruplex. The downregulation of luciferase expression via the presence of the 5'-UTR sequence of APC upstream of the luciferase reporter gene was confirmed (**Figure 32C**). The likely reasons for this difference could be: (i) the single nucleotide loop of G56 (**Figure 30D**) is very well protected from cleavage; and, (ii) the conditions found in the cell, specifically crowding or the presence of G-quadruplex binding proteins, might provide further stabilization of the G-quadruplex. Differences between *in vitro* and *in cellulo* results were also observed for the candidates with central loops of >69 nt (MDS1 and LRRC37A3). Even though in-line probing results showed patterns of G-quadruplex formation, no difference in luciferase activity was measured between the wt and G-tracts G/A-mutant constructs (data not shown), indicating either that cellular conditions are not favorable for the formation of G-quadruplexes with such long central loops or that they are not stable enough to affect translation significantly. In conclusion, the observed decreases in luciferase activity demonstrated that G-quadruplexes that include a long central loop up to 30-nt in length present inside the 5'-UTR are stable enough to negatively impact an essential biological process, in this case mRNA translation.

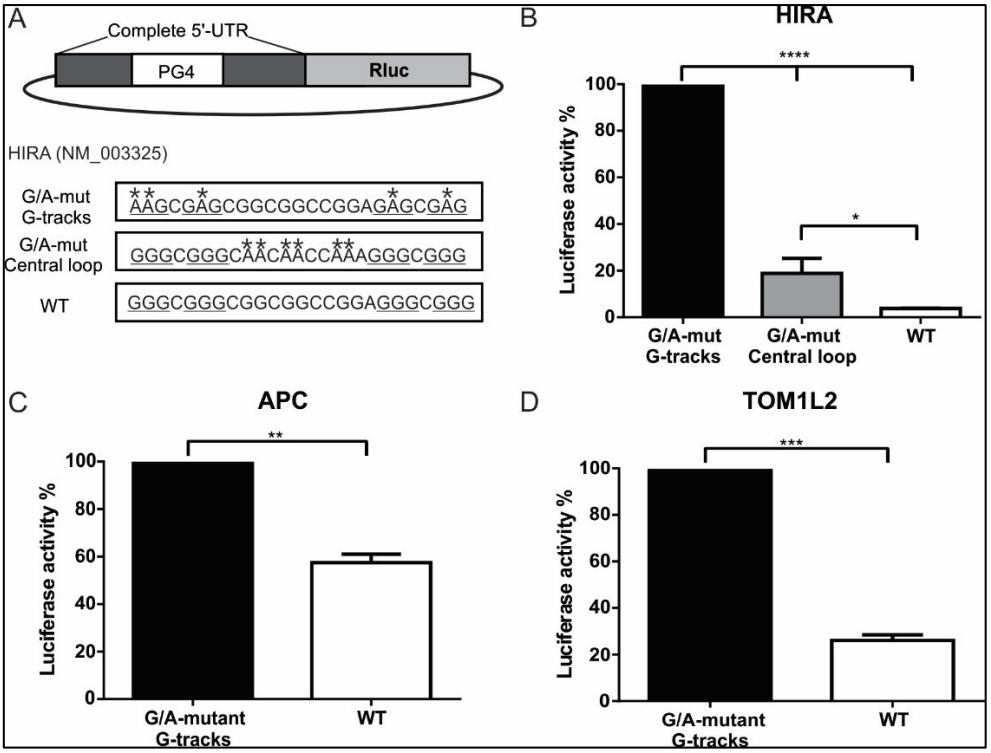



**Figure 32** – Effect of a G-quadruplex possessing a long central loop on luciferase activity.

(A) Schematic representation of the vector construction with the different sequences used for the HIRA candidate constructs. The PG4 region is boxed, the guanines involved in the G-tracks are underlined and nucleotides identified with an asterisk are the guanines that were mutated to adenines in the different G/A-mutant constructs. Values were first normalised by dividing the value of Rluc by the value of the control Firefly luciferase (Fluc). The percentage (%) of luciferase activity was set to 100% for all of the G-tracks G/A-mutant constructs. The luciferase activity values of the other construct were divided by the value of their corresponding Gtracks G/A-mutant construct and then multiplied by 100. (B) Luciferase activity of the different HIRA constructs each with a central loop of 11 nt. (C) APC possessing a central loop of 30 nt, and (D) TOM1L2 possessing a central loop of 32 nt. For these three examples, the wt constructs which can fold into a G-quadruplex reduced the luciferase activity. The results are the means of at least two independent experiments, and the error bars represent the standard deviations. Pvalue were calculated by unpaired Student t-test. (*) $P<0.05$ (**) $P<0.01$ (***) $P<0.001$ (****) $P<0.0001$.

## DISCUSSION

The results presented above confirm that potential RNA G-quadruplex forming sequences located in human 5'-UTRs harboring a long central loop (2–90 nt) and two single nucleotide distal loops are relatively common and might be physiologically relevant. Both *in vitro* and *in cellulo* data are in agreement with the earlier work of others (Bugaut et Balasubramanian, 2012 ; Guédin *et al.*, 2010 ; Pandey *et al.*, 2013) and question the legitimacy of the very strict G-quadruplex search algorithm that has been used in many studies, an algorithm which

considers loops only up to 7 nt long. Although this report changes the frontier of what might still be considered as putative G-quadruplex forming sequence, the approach used did not permit the elucidation of the upper limit of loop length consistent with G-quadruplex formation. The comprehensive bioinformatic search reported here has identified 1 453 5'-UTR PG4 sequences possessing a long central loop located on the complementary strand. In comparison, a similar survey published earlier by our laboratory using the above mentioned overly strict search algorithm limited to loops consisting of maximum 7 nt identified 7 198 PG4 sequences located on the complementary strand (Beaudoin et Perreault, 2010). If only PG4s with a central loop of ≥8 nt in length are considered, 1 232 additional PG4s possessing a long central loop went unnoticed by the previous limited search. This number represent a 17.11 % increase in newly identified PG4s. It is likely that additional search as for PG4s harboring either a long first or third loop accompanied by two single nucleotide long loops would further increase the number of therefore unidentified PG4s. However, thermodynamic data from previous biophysical studies carried out on artificial DNA sequences do not support the folding of G-quadruplexes with such loop arrangements (Guédin *et al.*, 2010). It is proper to note that bioinformatical approaches usually overestimate the actual number of G-quadruplexes present in the cell since they are restricted to sequence criteria only. Moreover, an analysis performed with a recently published scoring system used to identify RNA G-quadruplex folding (Beaudoin *et al.*, 2014) suggests that 40% of the PG4 candidates identified *in silico* are prone to fold into G-quadruplex structures based on a predicting value respecting both the ratio of consecutive guanines and the cytosine enrichment (data not shown).

This report investigated eight G-quadruplexes including long central loop. All of these RNA species were folded into G-quadruplex *in vitro* with the exception of APC*,* but only three out of the five tested *in cellulo* repressed translation*,* suggesting that less were formed in the cell. It is important to mention that the PG4 sequences that were chosen seemed to possess a high probability of folding into a G-quadruplex. For example, all of the selected candidates seemed to lack a Watson-Crick base-pair-based secondary structure stable enough to compete against the G-quadruplex structure. This was perhaps the most important criteria used, as the goal was to unambiguously demonstrate that some PG4 possessing a long central

loop were effectively folded. It is clear that, among the 5'-UTR PG4 sequences retrieved, there is a proportion of these sequence that do not fold into the G-quadruplex structure.

G-quadruplexes are known to be topologically extremely variable, and the folding of the structure is often driven by more complicated pathways which do not necessarily respect a simple two state equilibrium model between the folded and unfolded state, as was recently demonstrated for the human telomeric DNA sequence (Bian *et al.*, 2014). The final topology of the structure is usually influenced by a combination of different intrinsic and extrinsic factors, including the sequence of the molecule itself, the nature and concentrations of any monovalent ions, molecular crowding, the pH, and the temperature, among others. Unlike artificially designed sequences, which were primarily used in various biophysical studies in order to avoid the formation of unwanted folding possibilities, PG4s within biologically relevant regulatory regions such as the UTRs are very diverse in terms of G-tracts and loop lengths. This feature determines the variability in the number of stacked G-quartets and connecting loops. The presence of multiple G-quadruplex species in solution is one of the major problems complicating data evaluation in many biophysical approaches, including circular dichroism, NMR and UV-melting, where the resulting data often represents a mixture of different DNA G-quadruplex structures (Víglaský *et al.*, 2010). This report demonstrates that additional G-tracts located either in the loops or the regions flanking the predicted PG4s readily fold into a mixture of different G-quadruplex structures (see the candidates BAG1, HIRA and CTGLF6; **Figures 27-29**). In light of this finding, in-line probing appears to be the method of choice for assessing the complexity of all of the folding possibilities, which are then further reinforced by structural information. Among other advantages of in-line probing, the requirement of only trace amounts of RNA (<1 nM), which should favor intramolecular folding, and the ability to study short and as well as long RNA molecules under different salt conditions should be stressed. It is noteworthy that the folding of central loop sequences exceeding the length of 8 nt performed by RNAfold (Lorenz *et al.*, 2011) revealed that the vast majority of the sequences adopt a stem-loop secondary structure (data not shown). The coexistence of multiple G-quadruplex species, the exceptional length of some PG4s, and the very likely presence of an alternative structure in the central loop represents a limiting factor for well-established techniques such as circular dichroism and UV-melting. To avoid the limitations and data misinterpretation of the *in vitro* experiments,

the folding of some selected candidates was verified *in cellulo* by cloning the entire 5'-UTR containing the PG4 of interest upstream of the luciferase reporter gene. This approach successfully demonstrated the downregulation of luciferase expression for the wt sequences when compared to the mutated one for candidates with 11-, 30- and 32-nt long central loops. This strongly implies that G-quadruplexes with long loops might be stable enough to regulate gene expression on a cellular level.

This work demonstrates that it is possible to find G-quadruplexes possessing a long central loop in human 5'-UTRs. In addition, the folding of some interesting candidates possessing a central loop varying in length from 11 to 71 nt *in vitro* and 11 to 32 nt *in cellulo* has been confirmed. It is noteworthy that the presence of any extra G-tracts in the central loop provides additional folding pathways, resulting in the presence of multiple G-quadruplex species. The introduction of mutations that abolish the participation of these extra G-tracts in the central loop seems to be an effective way of regulating the folding of G-quadruplexes. The increased in-line cleavage of the nucleotides amid the guanine doublets in the central loop of the HIRA candidate indicates that G-quadruplexes with only two G-quartet layers might be in competition with the more stable one consisting of three G-quartets located within the same RNA molecule. The case of CTGLF6 provides proof that two G-quadruplexes arranged in tandem within one RNA molecule might coexist at the same time, and that the mutations can promote the folding of a particular structure. The *in vitro* folding of MDS1 and LRRC37A3, both of which possess exceptionally long central loops of 71- and 69-nt, respectively, defies the widely accepted definition of a G-quadruplex and calls for a revision of the previously established algorithm that considers only 7-nt- long loops. The existence of G-quadruplexes possessing long loops provides additional targets for drug design and new sites for protein–G-quadruplex interactions.

## MATERIAL AND METHODS

### Bioinformatics

The potential human G-quadruplex sequences used in this study were chosen from a 5'-UTR database derived from UTRdb and Transterm (Jacobs *et al.*, 2009; Mignone *et al.*, 2005). PG4s were identified using the program RNAmotif (Macke *et al.*, 2001) by describing an algorithm respecting the pattern $Gx\text{-}N_1\text{-}Gx\text{-}N_{2\text{-}90}\text{-}Gx\text{-}N_1\text{-}Gx$, were G stands for a guanine and

N for any nucleotide (A, U, C and G). The retrieved sequences were further analyzed using home written perl scripts, and were manually cured to obtain the database of PG4s possessing a long central loop provided in the Supplementary Information as an Excel sheet.

**RNA synthesis**

All sequences used in the *in vitro* experiments were synthesized by *in vitro* transcription using T7 RNA polymerase as described previously (Beaudoin et Perreault, 2010). Two overlapping oligonucleotides (2 mM each, Invitrogen) were annealed and a double-stranded DNA was obtained by filling in the gaps using purified Pfu DNA polymerase in the presence of 5% dimethyl sulfoxide (DMSO, Fisher). The double-stranded DNA sequence was then ethanol-precipitated. The resulting DNA templates contained the T7 RNA promoter sequence followed by the PG4 sequence. Transcription reactions were performed in a final volume of 100 µl using purified T7 RNA polymerase in the presence of RNase OUT (20 U, Invitrogen), pyrophosphatase (0.01 U, Roche Diagnostics) and 5 mM NTP in a buffer containing 80 mM HEPES-KOH pH 7.5, 25 mM $MgCl_2$, 2 mM spermidine and 40 mM DTT. The reactions were incubated for 2 h at 37°C, at which point they were treated with DNase RQ1 (Promega) for 20 min at 37°C. The RNA was then purified by phenol: chloroform extraction followed by an ethanol precipitation. RNA was fractionated by denaturing 10% polyacrylamide gel electrophoresis (8 M urea) (PAGE ; 19:1 acrylamide to bisacrylamide) using 45 mM Tris-borate pH 7.5, 1 mM EDTA solution as running buffer. After electrophoresis, the RNAs were visualized by UV shadowing and the bands corresponding to the correct size of the PG4s were excised from the gel and the transcripts eluted overnight at room temperature in buffer containing 1 mM EDTA, 0.1% SDS and 0.5 M ammonium acetate. The PG4s sequences were then ethanol-precipitated, dried and dissolved in water. The concentrations were determined by spectrometry at 260 nm using a NanoVue system (GE Healthcare).

**Radioactive 5'-end labelling**

In order to produce 5'-end-labeled RNA molecules, purified transcripts (50 pmol) were dephosphorylated at 37°C for 30 min by adding 5 U of antarctic phosphatase (New England BioLabs) in a final volume of 10 µl containing 50 mM Bis-propane pH 6.0, 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$ and 20 U RNase OUT (Invitrogen). The enzyme was inactivated by

incubation for 5 min at 65°C. The dephosphorylated RNAs (10 pmol) were 5'-end-radiolabeled using 7.5 U of T4 polynucleotide kinase (Promega) for 1 h at 37°C in the presence of 3.2 pmol of [$\gamma$-$^{32}$P]ATP (6000 Ci/mmol ; New England Nuclear). The reactions were stopped by the addition of two volumes of formamide dye buffer (95% formamide, 10 mM EDTA, 0.025% bromophenol blue and 0.025% xylene cyanol). The RNAs molecules were purified by 10% polyacrylamide 8 M urea gel electrophoresis. The bands corresponding to the 5'-end-labeled RNAs were then detected by autoradiography and the portions of gel containing the correct sizes were excised and recovered as described in the RNA synthesis section. The eluted and precipitated 5'-end-labeled transcripts were dissolved in 20 µl ultrapure water, and the final radioactivity was calculated using a Cerenkov counter (Bioscan QC-2000).

**In-line probing experiment**

Trace amounts of 5'-end-labeled RNA (50 000 cpm, <1 nM) were heated at 70°C for 5 min and then slow-cooled to room temperature over 1 h in buffer containing 20 mM lithium cacodylate pH 7.5 and 100 mM of either LiCl or KCl in a final volume of 10 µl. Thereafter, the final volume of each sample was adjusted to 20 µl such that the final concentrations were 30 mM lithium cacodylate pH 8.5, 20 mM $MgCl_2$ and 150 mM of either LiCl or KCl. The reactions were then incubated for 40 h at room temperature, at which point the RNA was ethanol-precipitated in presence of glycogen and then RNAs dissolved in 20 µL of formamide loading buffer (95% formamide and 10 mM EDTA, 0,025% bromophenol blue). For the alkaline hydrolysis ladder, 50 000 cpm of 5'-end-labeled wt RNA (<1 nM) were dissolved in water in a final volume of 5 µl, 1 µl of NaOH was added and the reaction incubated for 1 min at room temperature prior to being quenched by the addition of 3 µl of 1 M Tris-HCl pH 7.5. The RNA molecules were then ethanol-precipitated and dissolved in 20 µl of formamide loading buffer. For the RNAse T1 ladder, 50 000 cpm of 5'-end-labeled wt RNA (<1 nM) were dissolved in 9 µl of buffer containing 20 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$ and 100 mM LiCl. The mixture was incubated for 2 min at 37°C in the presence of 0.6 U of RNAse T1 (Roche Diagnostic), and then was quenched by the addition of 20 µl of formamide loading buffer. The radioactivities of both the in-line probing samples and the ladders were measured, using a Cerenkov counter (Bioscan QC-2000) and equal

amounts in terms of counts per minute for all samples were fractionated on denaturing (8 M urea) 10% polyacrylamide gels. The resulting gels were dried and the bands visualized by exposing them to a phosphoscreen (GE Healtcare) and then analysing it using a Typhoon Trio instrument (GE Healthcare).

**Data analysis**

In-line probing gels were analyzed using the Semi-Automated Footprinting Analysis (SAFA) software (Das *et al.*, 2005; Laederach *et al.*, 2008). The RNase T1 ladder lane was used as the "anchor" line, using the guanines as cleavage sites for the sequence reference in SAFA. The raw intensities of each band under different salt conditions were determined and exported into a text file. The file was then opened with the Excel program in order to produce a usable table. Subsequently, the intensity of each band in the lanes representing the favorable conditions in the presence of KCl was divided by the intensity of the corresponding band in the LiCl lane (the unfavorable condition). Each in-line probing experiment was performed in duplicate. The averages and standard deviations were calculated for the $K^+/Li^+$ ratios for each nucleotide. These values were used to generate bar graphs, plotting the $K^+/Li^+$ ratio on the *x*-axis and the nucleotide sequence on the *y*-axis.

**In cellulo luciferase assay**

The complete 5'-UTR sequences of the wt and various G/A-mutants of the HIRA, APC and TOM1L2 candidates flanked by *Nhe*I restriction site was synthesized *in vitro* via multiple-steps of PCR annealing and filling in of sets of overlapping oligonucleotides (Invitrogen). Complete 5'-UTR of both the wt and G/A-mutant constructs of MDS1 and LRRC37A3 flanked by *Nhe*1 restriction sites were obtained by custom gene synthesis (Biomatik). The list of the oligonucleotides and complete 5'-UTR sequences used are available in the Supplementary Information. The G/A mutations were the same as those in the *in vitro* constructs. The constructs were inserted upstream of the Renilla luciferase (Rluc) reporter gene in the *Nhe*1 restriction site of the pRL-TK vector (Promega) or the psiCHECK-2 vector for the HIRA constructs (Promega). All sequences were verified by DNA sequencing.

HEK 293 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, Wisent) supplemented with 10% foetal bovine serum (FBS, Wisent) and 1 mM sodium pyruvate (Wisent) at 37°C in a 5% $CO_2$ and 100% $H_2O$ atmosphere. Twenty-four hours pre-

transfection, 1.3 x $10^5$ cells were seeded in a 24-wells plate. The next day, either 450 ng of pRL-TK vector (Rluc) and 50 ng of pGL3-control vector (Firefly luciferase reporter, Fluc) or 25 ng of psiCHECK vector (containing both Rluc and Fluc reporter genes) and 475 ng of pUC19 carrier vector were transfected with 0.5 µL of Lipofectamine 2000 (Invitrogen) per well. Twenty-four hours later, the cells were lysed in passive lysis buffer (Promega) and the luciferase assays were performed following the Dual-luciferase Reporter Assay manufacturer's protocol (Promega) using the Glomax 20/20 luminometer. The Rluc value was normalised over the Fluc value. The percentage (%) of luciferase activity was then set to 100% for all of the G-tracts G/A-mutant constructs, while the luciferase activity values of the other constructs were divided by the value of their corresponding G-tracts G/A-mutant and multiplied by 100. The means and standard deviations were calculated from at least two independent experiments. Statistical significance was evaluated with an unpaired Student t-test using the GraphPad Prism 6.02 software.

## SUPPLEMENTAL MATERIAL

The database of the 1 453 PG4 sequences with long central loop is available as an Excel sheet (Supp Database.xlsx) at URL: https://rnajournal.cshlp.org/content/20/7/1129/suppl/DC1

## Annexe 3

**Table S1** Complete 5'-UTR RNA sequences used for in cellulo assays

**Table S2** List of DNA oligonucleotides used for synthesis of the complete 5'-UTR constructs

## ACKNOWLEDGMENTS

We thank Jean-Denis Beaudoin for the initial construction of the database of 5'-UTR PG4s possessing a long central loop. This work was supported by a grant from the Natural Sciences and Engineering Research Council (NSERC Canada, grant number 155219-07) to J.P.P. R.J. was the recipient of the CIHR Frederick Banting and Charles Best Canada Graduate Scholarship Master's Award. J.P.P. holds the Chaire de recherche de l'Université de Sherbrooke en Structure et Génomique de l'ARN and is a member of the Centre de Recherche Clinique Etienne-Le Bel.

# ARTICLE 4 – G-QUADRUPLEXES FORMATION IN THE 5'UTRS OF MRNAS ASSOCIATED WITH COLORECTAL CANCER PATHWAYS

**Auteurs de l'article** : Jodoin, Rachel et Perreault, Jean-Pierre

**Statut de l'article** : Publié dans PloS One (2018), vol. 13, no. 12, p. e0208363

**Avant-propos** : Rachel Jodoin a réalisé toutes les expériences *in vitro* et *in cellulo* présentées dans cet article et en a fait l'analyse des résultats. Le manuscrit a été rédigé par Rachel Jodoin et Jean-Pierre Perreault.

**Résumé :**

Les G-quadruplexes d'ARN (rG4) sont des structures secondaires non canoniques composées de séquences G-riches. Plusieurs structures rG4 situées dans les 5'UTR des ARNm agissent en tant que répresseurs de la traduction en raison de leur haute stabilité qui est considérée comme nuisible au *scanning* ribosomal. Malgré tout, il n'est pas connu si ce phénomène est particulier à quelques ARNm précis, ou s'il indique un mécanisme global de régulation, basé sur la reconnaissance de cette structure, permettant de coréguler des ARNm associés à une voie biologique commune. L'analyse de l'ontologie des ARNm possédant un motif rG4 prédit dans leurs 5'UTR a révélé un enrichissement pour les ARNm associés à la voie du cancer colorectal. Des outils bio-informatiques de prédictions de G4, ainsi que des validations expérimentales *in vitro* ont été utilisés afin de confirmer et de comparer le repliement des rG4s prédits des ARNm associés à des voies dérégulées lors du cancer colorectal. Le repliement rG4 a été confirmé pour la première fois pour 9 ARNm. Un effet répressif sur l'expression d'un gène rapporteur en lignées cellulaires colorectales cancéreuses a été démontré pour 3 candidats rG4. Ce travail met en lumière les lacunes de la prédiction des rG4 et l'importance essentielle de la caractérisation expérimentale afin d'identifier précisément le repliement rG4 ainsi que les possibles similarités partagées entre les rG4 surreprésentés dans des voies biologiques importantes.

**Abstract**

RNA G-quadruplexes (rG4) are stable non-canonical secondary structures composed of G-rich sequences. Many rG4 structures located in the 5'UTRs of mRNAs act as translation repressors due to their high stability which is thought to impede ribosomal scanning. That said, it is not known if these are mRNA-specific examples, or if they are indicative of a global expression regulation mechanism of the mRNAs involved in a common pathway based on structure folding recognition. Gene-ontology analysis of mRNAs bearing a predicted rG4 motif in their 5'UTRs revealed an enrichment for mRNAs associated with the colorectal cancer pathway. Bioinformatic tools for rG4 prediction, and experimental *in vitro* validations were used to confirm and compare the folding of the predicted rG4s of the mRNAs associated with dysregulated pathways in colorectal cancer. The rG4 folding was confirmed for the first time for 9 mRNAs. A repressive effect of 3 rG4 candidates on the expression of a reporter gene was also measured in colorectal cancer cell lines. This work highlights the fact that rG4 prediction is not yet accurate, and that experimental characterization is still essential in order to identify the precise rG4 folding sequences and the possible common features shared between the rG4 overrepresented in important biological pathways.

## INTRODUCTION

RNA G-quadruplexes (rG4) are non-canonical secondary structures based on the stacking of multiple g-quartets. A G-quartet is a coplanar array of four guanines (G) linked by Hoogsteen base-pairs and stabilized in its center by a monovalent cation, usually $K^+$. In the recent years, many rG4 located in the 5'UTRs of mRNAs have been described (Cammas et Millevoi, 2017 ; Fay *et al.*, 2017 ; Rouleau *et al.*, 2017b). To date, at least 35 examples of rG4 folding affecting expression levels *in cellulo* have been reported, including that of the well-known oncogene N-Ras (Kumari *et al.*, 2007). Based on a search for canonical motifs only, more than 2 000 human 5'UTRs were predicted to possess potential rG4 structures (PG4) (Beaudoin et Perreault, 2010 ; Huppert et Balasubramanian, 2005). Experimentally, a recent study using a next-generation sequencing technique called rG4-seq showed an enrichment of rG4 in the UTRs of mRNAs (Kwok *et al.*, 2016a). More specifically in the 5'UTR, 540 regions appeared to fold into the structure, many of these with sequence features that were divergent from the canonical description of an rG4 motif.

There is evidence of rG4 formation in the cell cytoplasm from experiments using both fluorescent antibodies and chemical probes specific for the structure (Biffi *et al.*, 2014a; Laguerre *et al.*, 2015). However, recent work using *in vivo* DMS-probing demonstrated that, in eukaryotes, rG4 are mostly unfolded (Guo et Bartel, 2016). This could indicate that the rG4 motifs identified in the transcriptome are either prevented from folding, are actively unfolded, or that the rG4 structures might be transient and folded only in specific regulatory mechanisms. Actually, some RNA-binding proteins are known to specifically recognize and bind rG4 structures [reviewed in (Fay *et al.*, 2017)]. Helicases such as DHX36 (Lattmann *et al.*, 2011), DDX21 (McRae *et al.*, 2017) and DHX9 (Chakraborty et Grosse, 2011) unfold the structure. Due to the high stability of the structure, rG4s in 5'UTRs were defined primarily as translational repressors that impaired ribosomal scanning (Bugaut et Balasubramanian, 2012). There are also a few examples of 5'UTR rG4s acting as translational enhancers, or as a part of the internal ribosome entry sites that are important for cap-independent translation (Agarwala *et al.*, 2013; Bonnal *et al.*, 2003; Morris *et al.*, 2010).

At the DNA level, G4 probable sequences are enriched in oncogenes and depleted in tumor suppressor genes (Eddy et Maizels, 2006). Recent work also demonstrated that G4 located in the promoters of DNA repair genes are folded and might play a role in the oxidative

stress response (Fleming *et al.*, 2018). At the RNA level, Kwok *et al.* observed an enrichment of rG4-seq detected rG4 in the mRNAs of genes involved in RNA processing, stability and transcription regulation (Kwok *et al.*, 2016a). Other than translation, mRNA-specific rG4s were also studied for their roles in post-transcriptional processes (Fay *et al.*, 2017; Millevoi *et al.*, 2012), including some that are important in diseases such as cancer or neuropathology (Cammas et Millevoi, 2017). Actually, G4 structures are considered as potential therapeutic targets and multiple efforts are driven toward the rational design of small-molecule ligands that would target the rG4 structure in specific mRNA transcripts (Parrotta *et al.*, 2014). All of this transcript-specific evidence points to the hypothesis that, similar to the DNA G4s located in promoters, rG4s located in the 5'UTRs might be structural motifs responsible for the co-regulation of the expression levels of mRNAs with different functions in order to regulate either global pathways or cellular responses.

The primary method for identifying potential G4 (PG4) is the presence of the consensus sequence motif. The canonical G4 were originally described as $G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$-$N_{1-7}$-$G_x$, where $x \geq 3$ (Huppert et Balasubramanian, 2005). The consecutive Gs form the four essential G-tracts that are linked by three series of any of the four nucleotides (N) that are called loops. However, extensive studies have now characterized rG4 folding for a broader array of motifs: the stacking of only two quartets (Liu *et al.*, 2002), or of more than three (Fratta *et al.*, 2012), the presence of loops longer than seven nucleotides (Bolduc *et al.*, 2016; Jodoin *et al.*, 2014), the presence of bulges in the G-tracts (Martadinata et Phan, 2014), or even completely different and unpredictable motifs such as the G-quadruplexes of the fluorescent RNA aptamers Spinach and Mango (Huang *et al.*, 2014; Trachman *et al.*, 2017). All of which renders the prediction of rG4 formation extremely difficult with a large spectrum of possible motifs. Moreover, the canonical motif in itself is not enough to result in rG4 formation, as the nucleotide context of the motif, for example the presence of C-rich sequences, can compete with rG4 formation and favor the formation of a double-stranded RNA structure (dsRNA) instead. This observation lead to the development of several G4 prediction tools that can measure G4 propensity, including measuring the competing nucleotide context (cG/cC score and G4H) (Beaudoin *et al.*, 2014; Bedrat *et al.*, 2016), the G4 homology by comparing to experimentally confirmed rG4 (G4NN) (Garant *et*

*al.*, 2017) and predicting the possible secondary structures in order to identify the most stable one (RNAfold) (Lorenz *et al.*, 2013).

Many biophysical techniques exist that can be used to both confirm the presence of and characterize rG4 folding, techniques such as a specific CD spectra signature, UV-absorbance and thermal denaturation (Weldon *et al.*, 2016). Dyes also provide fluorescence enhancement upon binding to specific G4 topologies (Nicoludis *et al.*, 2012 ; Renaud de la Faverie *et al.*, 2014). All of these techniques offer a global idea of the folding as either dsRNA or rG4, but none precisely define the nucleotides involved in either the base-pairs or the G-tracts of the rG4. In addition, these techniques can be used only with short nucleotide sequences. *In-line* probing is specific for RNA sequences and it was adapted to rG4 probing (Beaudoin *et al.*, 2013). Compared to the above-mentioned methods for studying the rG4 motif, it can be performed with longer sequences, which makes it useful in the competing nucleotide context, and it generates significantly more information about the flexibility of the individual residues in the structure.

As potential rG4 structures in 5'UTRs are pervasive and biologically relevant, whether or not the presence of rG4 in the 5'UTRs of different mRNAs could be related to their common regulation, or association with a similar pathway, were investigated using gene-ontology enrichment. As rG4 are difficult to predict based solely on the presence of the motif, the accessible tools for the prediction of rG4 were used and their results were compared to those of experiments using *in-line* probing and fluorescence enhancement assays that confirmed the folding *in vitro.* This permitted the observation of the rG4s features that are shared by the mRNAs of the same pathway ontology, as well as the *in cellulo* measurement of the rG4s effects of some candidates using gene-reporter expression assays. This work sheds light on both the remaining flaws of the rG4 prediction tools, and on the importance of the experimental characterization of individual rG4 in order to accurately identify them as they are possible mRNAs structural co-regulatory motifs.

## MATERIAL AND METHODS

### PG4 database

Databases of PG4 located in the 5'UTRs of mRNAs corresponding to the canonical motif $G_3$-$N_{1-7}$-$G_3$-$N_{1-7}$-$G_3$-$N_{1-7}$-$G_3$ are available from Beaudoin and Perreault (Beaudoin et

Perreault, 2010) and PG4 with a longer central loop $G_3$-$N_1$-$G_3$-$N_{1\text{-}20}$-$G_3$-$N_1$-$G_3$ from Jodoin *et al.* (Jodoin *et al.*, 2014). Briefly, the databases were built by retrieving all 5'UTR sequences from the database UTRdb (Mignone *et al.*, 2005). Python scripts were then used to search for the sequence motif and to identify its position.

**Bioinformatic methods**

The gene ontology enrichment analyses were performed using the DAVID bioinformatic resources version 6.7. A list of the 2 004 mRNAs' 5'UTRs containing at least 1 PG4 corresponding to either the canonical sequence or one with a central loop up to 20 nts in length were compared to the background of all human 5'UTR mRNAs (a total number of 31 654). Pathways were recovered using the KEGG orthology database (Kanehisa *et al.*, 2017). All of the candidates' mRNAs information (RefSeq number, UTRdb ID and Kegg orthology or AmiGO annotations (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2017)) are reported in **S3 Table in Annexe 4**.

Following the selection of the 26 PG4 candidates, the cG/cC (Beaudoin *et al.*, 2014), the G4Hunter (Bedrat *et al.*, 2016) and the G4NN scores (Garant *et al.*, 2017) were measured using the G4 screener webserver (Garant *et al.*, 2018). The thresholds selected for rG4 formation were 3.0, 0.9 and 0.5, respectively, so as to maximize both sensitivity and selectivity. For the RNA sequences used in the *in vitro* experiments, the scoring window was set to 200 nts in order to include the full-lengths of all sequences, and to give only one value for each score.

RNA secondary structure prediction of the sequences used for the *in vitro* experiments was performed using the RNAfold tool version 2.1.0 from the VIENNA RNA suite (Lorenz *et al.*, 2011) and changing the default parameters in order to add g-quadruplex predictions. The resulting most stable secondary-structures are represented by dot-bracket notation, and the guanines predicted to fold into rG4 are represented by the "+" symbol.

**Construction of RNA sequences**

The sequences tested *in vitro* consist of the PG4 in question surrounded by 15 to 50 nts of its natural 5' and 3' contexts. The G/A-mutants were designed to disrupt the G-tracts. Hence, each second G of a tract was mutated to A (typical examples: GGG were mutated to GAG, GGGG to GAGA or GAAG, etc.). The 17 nts T7 promoter sequence

(TAATACGACTCACTATA) were added for *in vitro* transcription purposes, followed by 2 or 3 Gs if they were not already present in the 5'UTR. The sequences of all of the oligonucleotides used are presented in **S4 Table in Annexe 4**. PCR templates for *in vitro* transcription were obtained by PCR filling of the 2 complementary oligonucleotides (IDT or Invitrogen) using purified PFU DNA polymerase (12 cycles of 1 min each at 95°C, 54°C and 72°C, followed by a final elongation at 72°C for 5 min) in buffer containing 0.2 mM dNTPs, 2 mM $MgSO_4$, 20 mM Tris-HCl pH 8.8, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 0.1% Triton-X-100 and 5 mM DMSO. The DNA template sizes were verified by agarose gel electrophoresis. The PG4 DNA templates were ethanol precipitated, dried and dissolved in 50 µL $H_2O$. *In vitro* T7 RNA transcription reactions were performed for 2 h at 37°C using 10 µg of purified T7 polymerase in a solution with 5 mM rNTPs, 0.01 U pyrophosphatase (Roche Diagnostics), 80 mM HEPES-KOH pH 7.5, 24 mM $MgCl_2$, 40 mM DTT and 2 mM spermidine. In order to remove the DNA template and to remove protein contaminants from the transcription reactions, a DNAse treatment (RQ1 DNAse, Promega) followed by phenol-chloroform extraction and ethanol precipitation were performed. The recovered RNA was separated on an 8% denaturing polyacrylamide gel (PAGE ; 19:1 ratio acrylamide to bisacrylamide, 8 M urea using 45 mM Tris-borate pH 7.5 and 1 mM EDTA solution as running buffer). The bands were visualized by UV-shadowing, the corresponding gel slices were cut out and the RNA eluted in elution buffer (1 mM EDTA, 0.1% SDS and 0.5 M ammonium acetate) and ethanol precipitated. RNA was dissolved in water and quantified using a Nanodrop Lite spectrophotometer (ThermoFisher scientific).

**5' end-labelling of RNA transcript**

RNA (50 pmol) was dephosphorylated in a 50 µL reaction using Antarctic phosphatase (1 U, New England Biolabs) using the manufacturer's protocol. The enzyme was inactivated by heating to 65°C for 7 min. Then, 10 pmol of dephosphorylated RNA was kept and $\gamma^{32}$P-ATP (2 µL ; 6 000 Ci (222 TBq)/mmol in 50 mM Tricine pH 7.6, PerkinElmer) was added along with 3 U of T4 Kinase (Promega) and the reaction was incubated for 1 h at 37°C. Labeled RNA was separated by denaturing PAGE as described previously. The RNA was detected by autoradiography. The correct RNA band was cut out of the gel and the RNA eluted, ethanol precipitated and dissolved in 30 µL of water. The radioactivity in counts-per-minute

(cpm) for each sample was measured using a single-well gamma particle counter (Bioscan QC-2000).

### *In-line* probing

*In-line* probing was performed as described previously (Beaudoin *et al.*, 2013). Briefly, the probing of each candidate was performed in duplicate from two different *in vitro* transcription reactions. Equal amounts, in terms of cpm (50 000 cpm), of 5'end labeled WT and G/A-mutant sequences were dissolved in folding buffer (10 µL ; 20 mM Li cacodylate pH 7.5 and 100 mM of either LiCl or KCl). Folding was performed by heating the RNA sample for 5 min à 70°C followed by a 1 h slow-cool to room temperature (RT). *In-line* 2X buffer (50 µL, 40 mM Li cacodylate pH 8.5, 40 mM $MgCl_2$ and 200 mM of either LiCl or KCl) and 40 µL $H_2O$ were then added so as to obtain a final volume of 100 µL. The samples were incubated for 40 h at RT in order to allow for self-cleavage to occur. The RNA was then ethanol-precipitated and dissolved in 20 µL of denaturing loading buffer (95% formamide, 10 mM EDTA, 0.025% xylene cyanol). Before separation on a 10% denaturing PAGE, the cpm of each sample was measured and the sample diluted if necessary so as to load an equal cpm amount of each. Both alkaline hydrolysis and RNase T1 sequence ladders where migrated alongside the samples. In order to obtain the alkaline hydrolysis ladder, a 5 µL solution of 50 000 cpm of either WT or G/A-mutant 5'end labeled RNA was treated with 2 µL of 2 N NaOH for 1 min at RT. The reaction was stopped by the addition of 3 µL of 1 M Tris-HCl pH 7.5. The RNA was then ethanol-precipitated and dissolved in 20 µl of denaturing loading buffer. In order to obtain the RNase T1 ladder, an 8 µL solution of 50 000 cpm of either WT or G/A-mutant 5'end labeled RNA was treated for 2 min at 37°C with 1 µL of RNAse T1 enzyme (0.6 U, Roche Diagnostic) and 1 µL of 10X buffer (200 mM Tris-HCL pH 7.5, 100 mM $MgCl_2$, 1 M LiCl). The reactions were stopped by the addition of 20 µL of denaturing loading buffer. The gel was migrated for 2 h at 60 Watts. The gel was then put on a Whatman paper and dried for 45 min at 80°C in a gel drier. The dried gel was exposed to a phosphorimager screen overnight, and cleavage pattern visualized using a Typhoon Trio imaging system (GE Healthcare).

**Quantification of the *in-line* probing data**

Quantification of gel band intensities was performed using the SAFA semi-automated software (Das *et al.*, 2005). Equal loading in the $K^+$ and $Li^+$ lanes of the *in-line* probing gels was first verified by quantifying the intensity of the top unresolved band. If there was less than a 15% difference the loading was considered as being equal, and the ratios of the intensities could be calculated. The intensities of each band in the $K^+$ conditions were divided by the intensities of each related band in the $Li^+$ conditions. The average $K^+/Li^+$ ratios of two independent experiments were represented on a graph for each nucleotide and for each condition (WT or G/A-mutant). The error bars represent the standard deviation. A $K^+/Li^+$ ratio threshold of 2 was set so as to conclude that a difference in nucleotide flexibility existed in the $K^+$ condition.

**NMM fluorescence assay**

The RNA WT and G/A-mutant sequences used for the NMM fluorescence assay were the same as those used for the *in-line* probing experiments. After *in vitro* transcription, the RNA was quantified using a Nanodrop Lite spectrophotometer (ThermoFisher scientific). RNA (200 pmol) was dissolved in 50 µL of the same folding buffer as was used in the *in-line* probing experiments. The RNA was heated to 70°C for 5 min, and then was slow-cooled to RT for 1 h. *In-line* 2X buffer (50 µL) was then added. NMM 0.5 mM (1 µL ; N-Methyl-Mesoporphyrin IX, NMM580, Frontier Scientific Inc., Logan, Utah,) was then added and the mix was incubated for 30 min at RT in the dark. Fluorescence spectrophotometry was performed using a Hitachi F-2500 fluorescence spectrophotometer with an excitation bandwidth of 399 nm, and the emission spectra was measured from 550 to 650 nm in a 10 mm quartz cuvette. The fluorescence units of the maximal peak at 605 nm was used to compare the $Li^+$ and $K^+$ conditions. The experiments were performed in triplicate with RNA sequences from three independent *in vitro* transcription reactions. The results are presented as the differences of the means of the $K^+$ and $Li^+$ fluorescence peaks at 605 nm.

**Cell cultures**

HEK293 cells (origin ATCC, CRL-1573) were cultivated in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% foetal bovine serum (FBS). HCT116 and HT-29 cells (origin ATCC, CCL-247 and ATCC, HTB-38, respectively) were cultivated in McCoy's

5A medium supplemented with 10% FBS. DLD-1 cells (origin ATCC, CCL-221) were cultivated in Roswell Park Memorial Institute 1640 medium (RPMI) supplemented with 10% FBS. In every case the incubations were performed in an incubator at 37°C with a 100% $H_2O$ and 5% $CO_2$ atmosphere. All cell culture reagents were purchased from Multicell and Wisent.

**Cloning and transfection**

Full length WT 5'UTR and G/A-mutant sequence constructs (the same mutations as were used in the *in vitro* experiment) with *NheI* restriction sites at both ends were generated by PCR filling of two complementary DNA oligonucleotides (Invitrogen) using the same protocol as described in the section: Construction of RNA sequences. The sequences and primers used for cloning are listed in the **S5 Table in Annexe 4**. Full length 5'UTR constructs with the *NheI* restriction sites were digested and ligated into the *Renilla* luciferase (Rluc) pRL-TK vector (Promega). The correct insertions and mutations were verified by DNA sequencing. The Firefly luciferase (Fluc) PGL3 vector (Promega) was used as a transfection control.

Twenty-four hours prior to transfection, HEK293 cells were seeded at 130 000 cells/well, while HCT116, HT-29 and DLD-1 cells were seeded at 190 000 cells/well in a 24-wells plates. Plasmid DNA (500 ng in total, 450 ng of the pRL-TK construction and 50 ng of pGL3 as control) were co-transfected using 0.5 µL/well of Lipofectamine 2000 as recommended by the manufacturer (Invitrogen) in the appropriate serum-free media for each cell type. Each candidate's WT and G/A-mutant constructions were transfected in triplicate, and each experiment was repeated at least twice for each candidate.

**Dual-luciferase assays**

Dual-luciferase assays were performed at RT using the manufacturer's protocol (Dual-luciferase reporter assay system, Promega). Briefly, cell lysis was performed 24 h post-transfection with 150 µL (HEK293) or 100 µL (HCT116, HT-29 and DLD-1) of passive lysis buffer 5X. A volume of 3 to 10 µl of cell lysate was used for the assay. The luciferase reagents (100 µL each) were added sequentially. Luminescence readings were performed with a Glomax 20/20 luminometer. The read integration times was 10 sec. The results are

presented as the means and standard deviations of the Rluc/Fluc ratios of at least two independent experiments.

**Statistical analyses**

Statistical analyses were performed using GraphPad Prism 7.03 (H.J. Motulsky, 2014).

## RESULTS AND DISCUSSION

### Potential 5'UTR rG4 folding motifs are enriched in annotated pathways

Numerous rG4 structures located in the 5'UTRs of individual mRNAs have been identified as translational repressors. However, it is not known if rG4 motifs are enriched in particular mRNA families or in certain cellular pathways. In order to answer this question, a gene ontology-enrichment (GO-enrichment) analysis was performed using the DAVID bioinformatic tool (Huang *et al.*, 2009). The previously described database of potential rG4 (PG4) developed by Beaudoin and Perreault was used (Beaudoin et Perreault, 2010). This database contains all of the 5'UTR sequences extracted from the UTRdb database (Mignone *et al.*, 2005) that contain the canonical motif $G_3$-$N_{1-7}$-$G_3$-$N_{1-7}$-$G_3$-$N_{1-7}$-$G_3$. Based on previous findings from our group indicating that rG4 with longer loops can fold with good stability, and can also affect gene expression (Jodoin *et al.*, 2014), the original database was conservatively extended to include 5'UTRs containing the motif $G_3$-$N_1$-$G_3$-$N_{1-20}$-$G_3$-$N_1$-$G_3$. This represents PG4s with a central loop of 1 to 20 nucleotides (nts), while the two other loops are limited to one nucleotide each. Further extension of the sequence motif search was avoided so as to limit the number of false rG4 predictions. The final list of 2 004 PG4 mRNAs was then compared to the background of all *homo sapiens* GO-annotated mRNAs in search of possible enrichment in certain biological pathways. The GO-annotations related to pathways were taken from the KEGG pathways database (Kanehisa *et al.*, 2017). The GO-annotations describe the molecular functions, the cellular components and the biological processes to which a gene and its corresponding mRNA transcripts are associated. Pathways annotations describe genes with various functions and cellular localizations that are commonly involved in higher order biological processes such as metabolic routes or the development of a disease. In the analysis performed here, an enrichment represents a higher proportion of the mRNAs in the PG4 mRNAs list being associated to a particular pathway as compared to the proportion of all mRNAs from the background that are associated with this

same pathway. The results of the GO-enrichment analysis are presented in **Table 6**. Seven pathways presented an enrichment of 1.8- to 2.5-fold for mRNAs with a PG4 motif located in the 5'UTR with significant *P*-values. Three of these were cancer-related pathways: acute myeloid leukemia, chronic myeloid leukemia and colorectal cancer. The signaling pathways of neurotrophin and insulin were enriched, as was glycerophospholipid metabolism and endocytosis. Interestingly, many PG4 containing mRNAs were in common between the different pathways such as BAD, MAP2K1 and PIK3R1, which were present in five of them. These mRNAs code for proteins involved in signalization, stress response and proliferation; molecular functions that are commonly altered in diverse cancers (Hanahan et Weinberg, 2011), explaining why they are retrieved in multiple cancer related pathways.

**Table 6** Gene ontology enrichment analysis

| Term | Count | % | *P*-Value | Genes | Fold Enrichment |
|---|---|---|---|---|---|
| hsa05221:Acute myeloid leukemia | 11 | 0.87 | 0.001 | CEBPA, NRAS, MAP2K1, STAT5A, MAPK3, STAT5B, RARA, BAD, PIK3R3, TCF7L1, PIK3R1 | 2.54 |
| hsa00564:Glycerophospho-lipid metabolism | 13 | 1.03 | 0.004 | GPD2, CPT1B, DGKQ, LYPLA1, CDS1, CDS2, DGKZ, ETNK2, PCYT1B, PPAP2A, AGPAT2, CHAT, AGPAT1 | 2.54 |
| hsa05220:Chronic myeloid leukemia | 13 | 1.03 | 0.008 | CTBP2, MAP2K1, STAT5A, STAT5B, SMAD4, BAD, ACVR1C, NRAS, CBLB, MAPK3, PIK3R3, CRK, PIK3R1 | 2.33 |
| hsa05210:Colorectal cancer | 14 | 1.11 | 0.010 | MAP2K1, SMAD4, SMAD2, FZD2, BAD, APPL1, TCF7L1, ACVR1C, FZD10, CASP9, BCL2, MAPK3, PIK3R3, PIK3R1 | 2.18 |
| hsa04722:Neurotrophin signaling pathway | 20 | 1.59 | 0.002 | YWHAZ, IRS2, MAP2K1, MAPK11, MAPKAPK2, BAD, NRAS, ATF4, PSEN1, PRDM4, MAP3K3, MAPK14, BCL2, MAPK3, SH2B3, NGFRAP1, SH2B1, PIK3R3, CRK, PIK3R1 | 2.13 |
| hsa04910:Insulin signaling pathway | 21 | 1.67 | 0.002 | IRS2, MAP2K1, EXOC7, SOCS3, FLOT1, RHOQ, BAD, PPP1CC, PPP1CB, PRKAR2B, NRAS, PPP1CA, CBLB, PDPK1, INPP5K, MAPK3, PRKACA, PIK3R3, TRIP10, CRK, PIK3R1 | 2.05 |
| hsa04144:Endocytosis | 25 | 1.99 | 0.005 | FGFR3, CHMP4B, CHMP6, ADRBK2, ARF6, ACVR1C, HSPA1L, RNF3, HSPA6, NEDD4L, IQSEC2, GIT1, PARD6A, EPN3, VPS45, RAB11FIP4, RAB11FIP5, CBLB, PSD, AP2A1, ARRB1, ACAP2, SMURF1, PARD6G, PIP4K2B | 1.80 |

The biological significance of the GO-enrichment depends on the accurate prediction of rG4 formation in the mRNAs. It is probable that some PG4 of the initial 2 004 mRNAs sequences used for the GO-enrichment analysis are false positive predictions. However, this list represented the best starting point with which to consider pathway enrichment for the global amount of 5'UTR PG4s. Refinement of the rG4 prediction using different tools, and folding evaluation of the PG4 sequences of the selected enriched pathways, were the next steps towards the validation of the initial results.

**Dysregulated colorectal cancer pathways include mRNAs with PG4s**

The colorectal cancer pathway was selected for the continuation of the investigation of the importance of rG4 motifs in mRNA expression regulation because of the important biological incidence of this cancer, and because the molecular aspects of its dysregulated pathways are well-characterized. Furthermore, of the 14 mRNAs containing PG4 located in the 5'UTR that are enriched in this pathway, 6 had been previously studied for rG4 formation and 4 were already known to adopt an rG4 conformation. These latter 4 were BCL-2 (Shahid *et al.*, 2010), FZD2 (Beaudoin et Perreault, 2010), ACVR1C and MAPK3 (Beaudoin *et al.*, 2014). These results provided confidence that the GO-enrichment observed in this pathway was not biased by the presence of a high number of mRNAs with false rG4 predictions.

However, the KEGG's list of the mRNAs associated with the colorectal cancer pathway does not include all of the mRNAs that are known to be dysregulated in this cancer type. In order to correct for the incomplete annotation, and to increase the number of candidates, the list was extended. To do so, the initial list of 2 004 mRNAs positive for the presence of PG4 motifs was re-analysed. The previous study of hundreds of colorectal cancer tumors analysed by the Cancer Genome Atlas Consortium defined four predominant dysregulated pathways: that are the WNT, TGF-β and PI3-Kinase signalling pathways, and the proliferation and apoptosis defects (Cancer Genome Atlas Network, 2012). Twelve mRNAs from the PG4 database were thus recovered based on their more specific GO-annotations, related to one or more of the four colorectal cancer dysregulated pathways mentioned above (the GO-annotations of all candidates are listed in **S3 Table in Annexe 4**).

Furthermore, one candidate that was not present in the initial PG4 mRNA list for the GO-analysis, because it differed from the 1 to 20 nts central loop pattern, was also added. The APC candidate has a predicted central loop of 30 nts. The formation of an rG4 by this candidate had been previously confirmed by *in-line* probing (Jodoin *et al.*, 2014). Well-known to be mutated and important in colorectal cancer tumorigenesis (Morin *et al.*, 1997), APC was added as a positive control for both rG4 formation and possession of a role in colorectal cancer. **Table 7** presents the list of the 26 5'UTR PG4 candidates selected for further prediction and evaluation, regrouped by their associated mRNAs' pathways.

**Table 7** List of PG4 located in the 5'UTRs of mRNAs that are associated with colorectal cancer, their prediction of rG4 formation and their probing results.

| *Pathways* | *Candidates* | *rG4 predictions*[1] | | | | *In vitro probing* | | *rG4 formation*[4] |
|---|---|---|---|---|---|---|---|---|
| | | cG/cC | G4H | G4NN | RNA fold | *In-line* probing[2] | NMM fluorescence[3] | |
| *WNT* | APC | 6.21 | 1.00 | 0.86 | dsRNA | Yes | 32.2 | + |
| | BCL-9L | 4.37 | 0.86 | 0.97 | rG4 | Yes | 63.3 | + |
| | FZD10 | 5.04 | 1.21 | 0.84 | rG4 | Yes | 61.7 | + |
| | FZD2 | 13.3 | 1.63 | 0.97 | rG4 | Yes | 67.5 | + |
| | TCF7L1 | 1.21 | 0.23 | 0.05 | rG4 | No | 4.30 | - |
| *Apoptosis* | AIFM2 | 4.33 | 0.89 | 0.35 | rG4 | Yes | 49.3 | + |
| | APPL1 | 1.92 | 0.46 | 0.20 | dsRNA | Yes | 46.1 | + |
| | BAD | 2.51 | 0.60 | 0.24 | dsRNA | Yes | 26.8 | + |
| | BAG-1 | 4.04 | 0.83 | 0.62 | rG4 | Yes | 82.0 | + |
| | BAG-5 | 2.28 | 0.60 | 0.35 | dsRNA | No | 31.4 | - |
| | BCL-2 | 1.93 | 0.44 | 0.05 | dsRNA | Yes | 52.9 | + |
| | BOK | 2.43 | 0.67 | 0.39 | rG4 | No | 14.1 | - |
| | CASP6 | 3.07 | 0.75 | 0.34 | dsRNA | No | 11.9 | - |
| | CASP8AP2 | 3.57 | 0.79 | 0.77 | rG4 | Yes | 70.8 | + |
| | CASP9 | 3.68 | 1.06 | 0.63 | rG4 | No | 51.9 | - |
| *TGF-β* | ACVR1C | 3.01 | 0.61 | 0.51 | dsRNA | Yes | 71.2 | + |
| | BMPR1A | 7.00 | 1.12 | 0.89 | rG4 | Yes | 87.6 | + |
| | SMAD2 | 1.59 | 0.29 | 0.03 | dsRNA | No | 35.2 | - |
| | SMAD4 #1 | 2.54 | 0.70 | 0.31 | rG4 | No | 10.8 | - |
| | SMAD4 #2 | 3.07 | 0.82 | 0.47 | dsRNA | No | 12.2 | - |
| | SMAD7 | 0.73 | -0.05 | 0.02 | dsRNA | No | 0.30 | - |
| | SMURF1 | 1.86 | 0.41 | 0.21 | rG4 | No | 3.80 | - |
| *PI3-Kinase* | MAP2K1 | 2.33 | 0.49 | 0.16 | rG4 | Yes | 81.1 | + |
| | MAPK3 | 9.00 | 1.63 | 0.97 | rG4 | Yes | 48.1 | + |
| | PIK3R1 | 2.97 | 0.83 | 0.43 | rG4 | Yes | 35.6 | + |
| | PIK3R3 | 2.13 | 0.52 | 0.21 | rG4 | No | 5.10 | - |

*1. Thresholds for a positive rG4 prediction are ≥3.0 for cG/cC ; ≥0.9 for G4H; and, ≥0.5 for G4NN*

*2. Based on the cleavage pattern, "Yes" represents sequences with a $K^+/Li^+$ ratio of cleavage equal to or superior to the threshold of 2 for the nucleotides located between tracts of guanines that is characteristic of rG4 folding. "No" represents either a $K^+/Li^+$ ratio inferior to the threshold, or a higher ratio that is either inconsistent or insufficient for rG4 folding.*

*3. Difference of the $K^+$ and $Li^+$ fluorescence emission peaks at 605 nm for the WT sequence.*

*4. Assignment of rG4 formation, "+" represents sequence positive for rG4 folding, "-" represents sequence negative for rG4 folding based on the two in vitro probing assays.*

***In silico* predictions of rG4 formation vary between different tools**

Previous work on rG4 have shown that their prediction based only on the presence of the sequence motif is prone to yielding many false positives (Beaudoin *et al.*, 2014 ; Beaudoin et Perreault, 2010 ; Bedrat *et al.*, 2016). Many factors can influence rG4 folding. For example, the presence of multiple cytosines (C) in the vicinity of the potential rG4 can compete with G-tract formation, resulting in G-C base pairs and folding into dsRNA instead of rG4. Therefore, a more detailed prediction of rG4 formation using the available bioinformatic tools was performed. The cG/cC and G4H scores were developed for RNA and DNA sequences, respectively (Beaudoin *et al.*, 2014; Bedrat *et al.*, 2016). They both use a similar window screening of sequences, but a different calculation method, to account for the possible unfavorable C-rich nucleotide context surrounding the potential rG4 motif. A third tool, the G4NN score (Garant *et al.*, 2017), was recently developed to measure sequence homology to experimentally-confirmed positive and negative rG4 sequences from the G4RNA database (Garant *et al.*, 2015). Based on the sensitivity and specificity analyses of these previous studies which compared multiple PG4 sequences, the optimized thresholds for G4 prediction were set to 3.0, 0.9 and 0.5 for the cG/cC, G4H and G4NN scores, respectively. Finally, Lorenz *et al.* (Lorenz *et al.*, 2013) developed an energy model for rG4 folding that was included in the RNAfold algorithm and allowed the comparison of the minimum free energies of the ensemble of dsRNA secondary structures versus that of the rG4 structure in order to identify the most stable one. Dot-and-bracket notation of the free or base-paired nucleotides of the secondary structure prediction was modified in order to add another symbol (+) indicating which Gs are involved in the G-tracts of the rG4. It was thus also used for the rG4 prediction for the set of 26 candidates.

The sequences selected for both the scoring and the secondary structure predictions were the PG4 sequence motifs obtained from the initial database. The sequence motifs ranged from 17 to 56 nts (3 overlapping PG4s) in size to obtain an average length of 27 nts. To account for the possible competitive secondary structure, a surrounding nucleotide context on either side (5’ and 3’) from the original 5’UTR was added. Based on our previous work (Beaudoin *et al.*, 2014 ; Beaudoin et Perreault, 2010), a nucleotide context ranging from 15 to 50 nts was ideal. The size of the added context for each of the 26 PG4 sequence depended on multiple factors. First, by the length of the 5’UTR, when possible the full 5’UTR was

selected. Second, the size of the context was affected by the position of the PG4 motif inside the 5'UTR. The context was shorter if the PG4 was close to the 5'or 3'extremity of the UTR. The sequence context from the coding region of the transcript was never included. Third, in order to perform further *in vitro* folding evaluations, constraints on the maximal sequence length were considered. *In-line* probing can be performed on RNA sequences up to 150 nts long. Furthermore, RNA synthesis using T7 polymerase-driven *in vitro* transcription is favored by the presence of multiple guanines at the 5' extremity, so the context size was selected to include natural 5'UTR context starting with multiple Gs when possible. Otherwise, 1 to 3Gs were added at the 5'-extremity. Considering all these technical limitations and each 5'UTR specificities, the resulting nucleotide context was of 4 to 60 nts in length, (average of 33 nts), and the resulting PG4 sequences studied were of 50 to 149 nts in length, with an average of 93 nts. The selected 26 PG4s sequences are available **in S1 Table in Annexe 4**, and the predictions of their formation by the four tools are presented in **Table 7**. The scores were calculated using one window covering the entire sequence.

Overall, the predictions vary significantly from one tool to the next. The four predictors gave identical predictions of rG4 formation, or dsRNA formation, for only approximately one-half (i.e. 12) of the candidates. G4H has the lowest rG4 prediction numbers, with 6, and RNAfold has the highest with 16 predicted positive candidates out of the total of 26 (**S2 Table in Annexe 4**). These divergent predictions result from the different stringencies of the tools in question.

### *In vitro* confirmation of rG4 folding

The rG4 predictors are a good starting point with which to identify strong PG4 candidates, but their accuracy can only be determined following experimental validation of the rG4 folding. To permit this evaluation, *in-line* probing cartography was performed on the same PG4 sequences and context used for prediction. The theoretical basis of this cartography is the self-cleaving potential of an RNA strand. The self-cleavage occurs when the flexible regions of the RNA strand adopt an "*in-line* conformation" between the 2'-hydroxyl group of the nucleotide's ribose moiety and the phosphate group of the backbone. The secondary structure is inferred from both the cleaved pattern of the flexible nucleotides and the protected pattern of the base-paired nucleotides. This technique presents multiple advantages for rG4 probing that have been described previously (Beaudoin *et al.*, 2013). For the 26 colorectal

PG4 candidates, the WT sequence was compared with its corresponding G/A-mutant sequence in which at least one G of each predicted G-tract was mutated to adenine (A) in order to abolish all folding of the potential rG4. A second negative control was performed using lithium ($Li^+$) instead of potassium ($K^+$) in the buffer. $K^+$ is essential for the stabilization of the G-quartets. The use of $Li^+$ offers a similar ionic strength in solution, that is unfavorable for rG4 stabilization, but do not affect any other dsRNA base-pairing formation. During incubation, self-cleavage of the RNA occurs in the flexible nucleotide regions of the folded RNA strands. Thus, in the case of an rG4 structure formation, only in the presence of $K^+$, the increase of cleavage will occur primarily for the nucleotides becoming single-stranded that are located in the loops and the regions immediately upstream and downstream of the rG4. Meanwhile, the G-tracts remain protected.

A representative example of *in-line* probing is shown in **Figure 33** for the Apoptosis related 5'UTR PG4 of the CASP8AP2 mRNA. Following radioactive labelling of the 5'extremity of the *in vitro* transcribed RNA sequence bearing the PG4 motifs and their surrounding nucleotide contexts (**Figure 33A**), the RNA was heat denatured and allowed to fold by slow cooling. The RNA was then incubated at room temperature for 40 h in the presence of either $Li^+$ or $K^+$ in order to allow self-cleavage to occur. The RNA was then separated on a denaturing PAGE gel along with the respective alkaline hydrolysis and RNAse T1 sequence ladders so as to be able to identify each nucleotide (**Figure 33B**). After exposition to a phosphor imaging screen, the density level of each band was measured and compared using the $K^+/Li^+$ cleavage ratio. An arbitrary 2-fold threshold was used to label a nucleotide as being flexible. The $K^+/Li^+$ ratios of the WT and G/A-mutant sequences of CASP8AP2 are presented in **Figure 33C**. Nucleotides flexible in the presence of $K^+$, and thus specific to the rG4 favorable condition, as well as protected G-tracts, are shown on **Figure 33A.** For the CASP8AP2 candidate, the cleavage pattern of the nucleotides in between G-tracts is representative of an rG4 folding. The presence of more than 4 protected G-tracts indicates the possibility of multiple co-existing rG4 formations using different combinations of loops and G-tracts. The "+" symbols below the nucleotides in positions 23-24-25, 28-29-30, 33-34-35 and 37-38-39 in **Figure 33A** indicate the G-tracts of the rG4 secondary structure predicted by the RNAfold algorithm. Because the guanine located at position 39 shows a high $K^+/Li^+$ ratio, it was considered as being flexible; hence, it is unlikely

that this series of Gs is part of the rG4. Thus, RNAfold positively predicted rG4 formation in the CASP8AP2 sequence, but at the incorrect G-tracts positions.

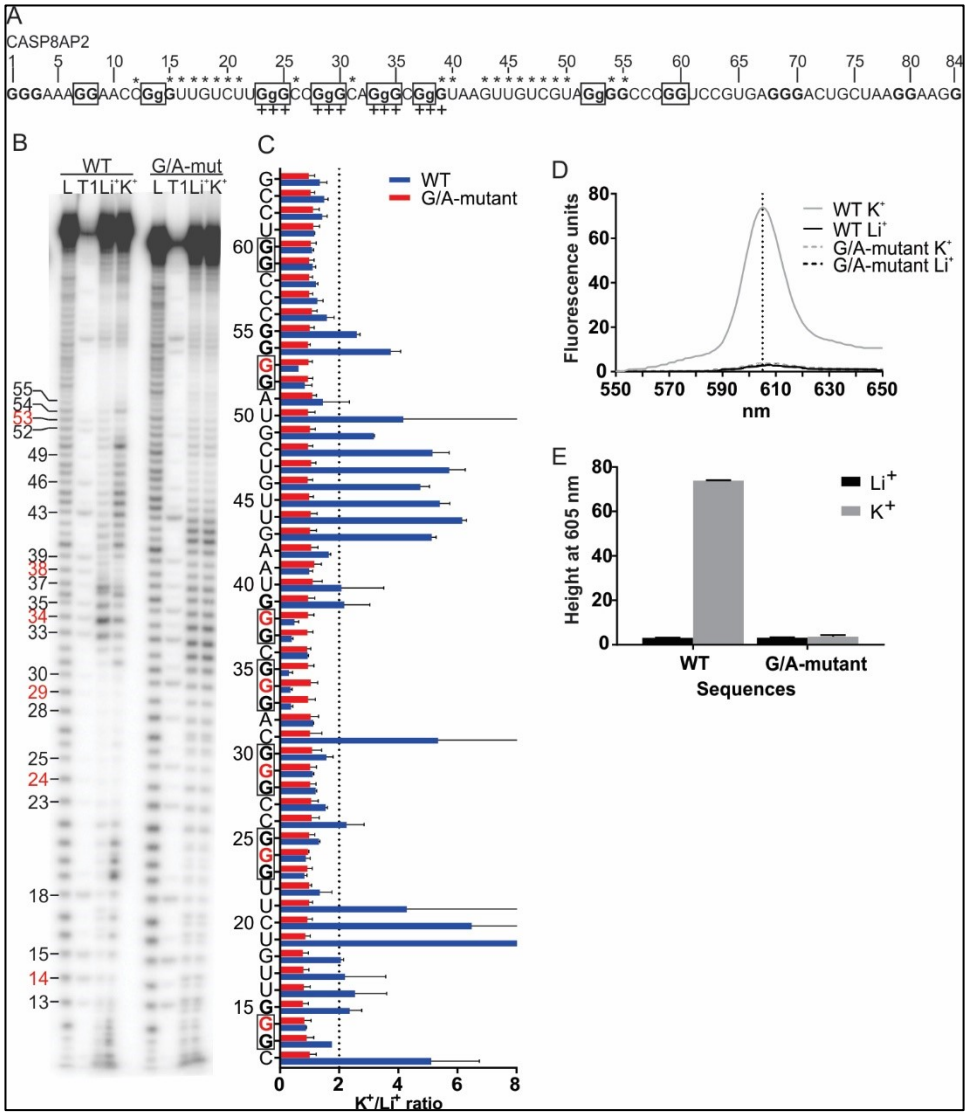



**Figure 33** – In vitro probing results for the candidate PG4 CASP8AP2.

A) Sequence, in the 5' to 3' orientation, of the PG4 with the G-tracts in bold and the G mutated to A in lower-case. Asterisks over the sequence indicate nucleotides for which the $K^+/Li^+$ cleavage ratio is higher than the threshold of 2. The "+" symbols indicate the Gs involved in G-tract formation as predicted by the RNAfold algorithm. The boxed G-tracts are those involved in G-quadruplex formation based on the *in-line* probing results. B) Representative phosphorimaging of a CASP8AP2 *in-line* probing denaturing PAGE gel. The alkaline hydrolysis ladder (L) and RNAse T1 ladder (T1) indicate the positions of every nucleotide and every guanine, respectively. Guanine numbering positions are indicated on the left, and the positions in red are those mutated to A in the G/A-mutant. C) $K^+/Li^+$ quantification of the *in-line* probing band intensities for each nucleotide for both the WT sequence (blue) and the G/A-mutant sequence (red). Each bar represents the mean of 2 independent experiments, and the error bars represent the standard deviations. The $K^+/Li^+$ cleavage ratio threshold

of 2 is indicated by the dotted line. D) Fluorescence emission curves of the WT and the G/A-mutant RNA sequences of the CASP8AP2 candidate in the presence of NMM after excitation at 399 nm. The full line represents the WT, the dotted line represents the G/A-mutant, the gray line indicate the presence of $K^+$ and the black line the presence of $Li^+$. Each curve is the mean of 3 independent experiments. The vertical dotted line indicates the 605 nm peak expected when NMM is bound to quadruplex RNA. E) Fluorescence emission peaks observed at 605 nm under the different conditions: Black, $Li^+$; Gray, $K^+$. Each bar represents the mean of 3 independent experiments, and the error bars are the standard deviations.

**rG4 formation was confirmed by an NMM fluorescence assay.**

In order to further validate the claim of rG4 formation in the CASP8AP2 candidate, a second *in vitro* supporting technique was used. N-Methyl Mesoporphyrin IX (NMM) is a ligand that had previously been shown to specifically bind to DNA G4 with a parallel topology (Nicoludis *et al.*, 2012). The ligand by itself emits a very low fluorescence, but upon binding to a parallel G4, its fluorescence can be increased from 2- to 10-fold. As RNA G4 mostly adopt parallel topology because of the *anti*-conformation of the ribose moiety (Tang et Shafer, 2006), the fluorescent enhancement of NMM in the presence of $K^+$ and WT sequences, as compared to that in presence of $Li^+$ and G/A-mutant sequences, was used as confirmation of the rG4 folding of the candidates. **Figure 33D** presents the fluorescent emission curves of NMM after excitation at 399 nm following a 30 min incubation with either the WT or the G/A-mutant sequences of CASP8AP2 under the various $Li^+$ and $K^+$ conditions. The characteristic peak at 605 nm for parallel G4 binding was observed. The measured fluorescence emission of the 605 nm peak is presented in **Figure 33E,** with enhancement only being observed in the rG4-prone WT and $K^+$ conditions. This result confirms the rG4 folding of the CASP8AP2 candidate observed by *in-line* probing.

**Approximately half of the 5'UTR PG4 sequences do fold *in vitro* into an rG4**

*In-line* probing of 19 candidates was performed, and the results were supplemented with the already available *in-line* probing data of the candidates APC from one study (Jodoin *et al.*, 2014) and FZD2, TCF7L1, ACVR1C, SMAD2, SMAD7 and MAPK3 from another (Beaudoin *et al.*, 2014). Representative *in-line* gels with quantifications from two independent experiments for all of the candidates are available in S1 Fig. The NMM fluorescence assay was performed for all 26 PG4 sequences. The results are presented in **S2 Fig. in Annexe 4.** The results of both *in vitro* techniques are summarized in **Table 7**. The rG4 formation was confirmed for 15 candidates, 9 of them for the first time (BCL-9L,

FZD10, AIFM2, APPL1, BAD, CASP8AP2, BMPR1A, MAP2K1 and PIK3R1). NMM fluorescence assays confirmed the previous conclusions of rG4 folding of APC, FZD2, BAG-1, BCL-2, ACVR1C and MAPK3, and the dsRNA folding of TCF7L1, SMAD2 and SMAD7. The dsRNA folding of the SMURF1 candidate observed here both by *in-line* probing and NMM fluorescence assay is different from that of a prior CD study (Mirihana Arachchilage *et al.*, 2014) which concluded that there was rG4 formation. In this prior study, only the region corresponding to positions 49 to 66 of the 100 nts sequence probed here was used. The presence of a biologically relevant competitive nucleotide context in the present study might explain the discrepancy between the 2 studies. The folding of a dsRNA structure was also observed for BAG-5, BOK, CASP6, CASP9, the two sequences from SMAD4 and PIK3R3. **Figure 34** presents the sequences probed for all of the 5'UTR PG4s of the mRNAs, classified by their associated pathway. The flexible nucleotides and protected G-tracts are identified. The PG4 sequences that were assigned positive for rG4 formation are presented with boxed G-tracts.

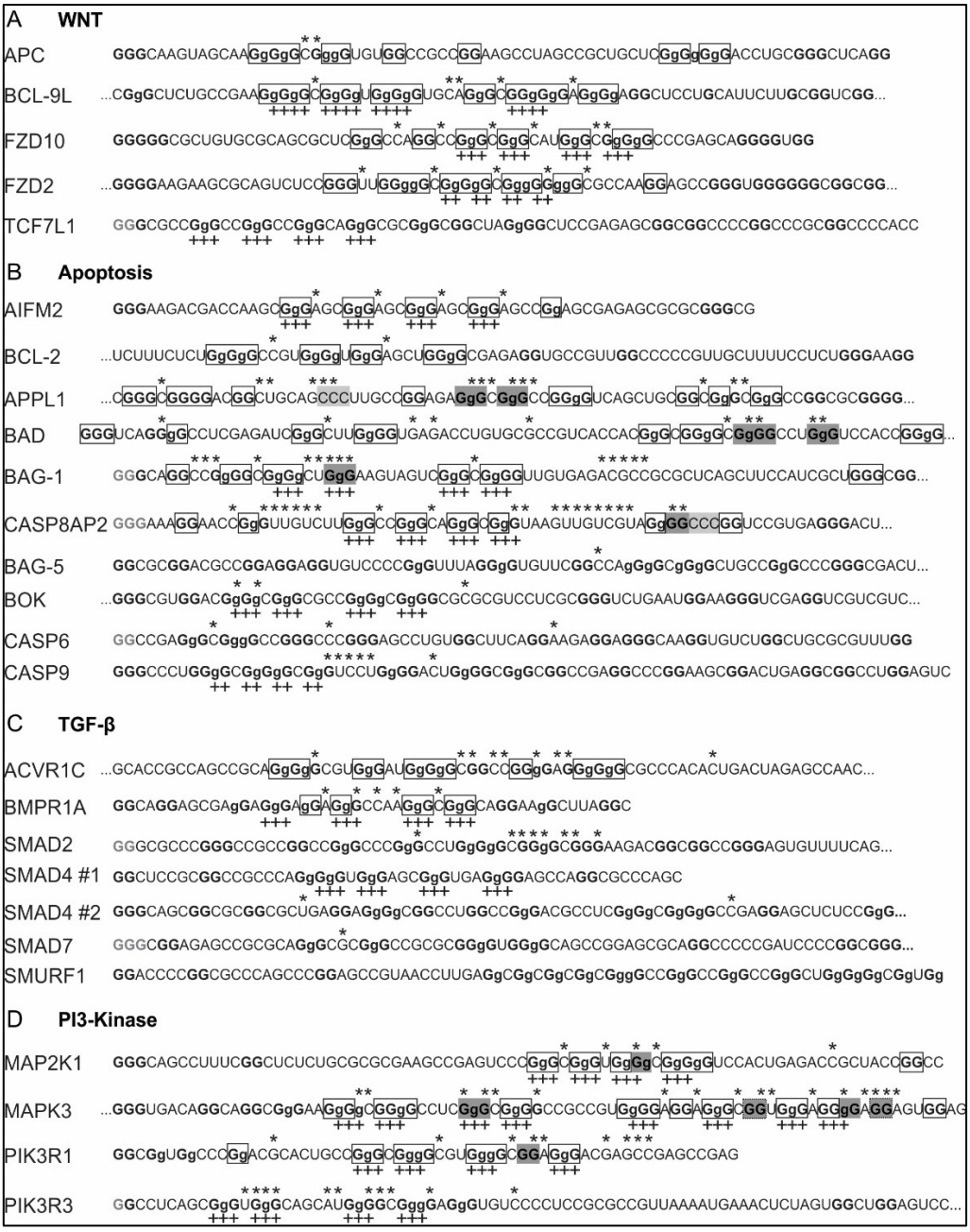



**Figure 34** – In line probing results for all candidates classified by pathway.

A) WNT; B) Apoptosis; C) TGF-β; and, D) PI3-Kinase. The sequences are in the 5' to 3' orientation, the series of 2 or more consecutive Gs are in bold and the G mutated to A in the mutants are in lower-case. The presence of three dots at either the beginning or the end of a sequence indicates that the RNA sequence actually tested was longer, but could not be presented fully here due to space limitations. The gray guanines at the 5' extremity were artificially added in order to optimize the *in vitro* transcription reactions. The full sequences tested are listed in **S1 Table in Annexe 4**. The asterisks over the nucleotides indicate a $K^+/Li^+$ cleavage ratio higher than the threshold of 2. The "+" symbols underneath the sequence indicate the Gs involved in G-tract formation as predicted by the RNAfold algorithm. The boxed G-tracts in white are those involved in G-quadruplex formation based on both the *in-line* probing and the NMM fluorescence assay results. The hatched boxes are the

G-tracts alternatively involved in G-quadruplex formation depending on the different G to A mutations (see MAPK3 in S1 Fig, panel B). Series of 2 or more consecutives Gs present in the loops are highlighted in dark gray, and series of 3 consecutive Cs in the loops are highlighted in pale gray.

Overall, the two experimental techniques used to evaluate rG4 or dsRNA folding were in good agreement. The *in-line* probing pattern of the cleavage representative of rG4 folding under $K^+$ conditions was generally associated with a high $K^+$ versus $Li^+$ difference in the 605 nm fluorescence peak enhancement (**Table 7**). The average difference value of the NMM 605 nm fluorescence peak was 16.5 ± 16.2 (mean ± standard deviation) for dsRNA and 58.4 ± 18.8 for rG4 folding. Because of the extent of variation in the NMM fluorescence, when the difference between the $K^+$ and $Li^+$ fluorescence emission peaks was intermediate (i.e. ~ 30) preponderance was given to the *in-line* probing results in the final assignment of rG4 (+) or dsRNA (-) structure formation. For examples, the BAD candidate presented an rG4 folding pattern using its *in-line* probing results, but a modest fluorescent enhancement and was assigned positive for rG4 folding. On the other hand, the CASP9 candidate did not presented a convincing rG4 pattern using *in-line* probing despite high fluorescence in presence of NMM and was assigned negative. However, it is not excluded that sequences with apparent contradictory results could adopt either folding types in different conditions than the ones assessed here. Variation in the fluorescence enhancement from one rG4 candidate to another could be explained by different rG4 features, such as both the G-tracts' numbers and sizes and the loops' sizes and compositions. The variation could also be explained by the proportion of RNA strands that folded as an rG4 versus as a dsRNA (e.g. the equilibrium between the competing structures) differing for each candidate. The results of the *in vitro* confirmation of rG4 folding demonstrate that many sequences possessing the consensus PG4 sequence motif prefer to adopt a dsRNA structure. This result is similar to that of a previous study of PG4 sequences (Beaudoin et Perreault, 2010). The rG4 prediction based on motif search only is unreliable. Thus, it is essential to confirm the folding with multiple reliable experimental techniques, and to use different prediction tools that take into consideration the competing nucleotide context.

**G4NN is the most accurate predictor of in vitro rG4 formation for this set of PG4**

With both sets of experimental results in hand, confirming or not rG4 formation, it is possible to compare the accuracy of the rG4 prediction tools. **S2 Table in Annexe 4** presents the

numbers of all true positive and negative predictions for each of the three scores, as well as for the RNAfold algorithm. These values were used to calculate the sensitivity and specificity of each predictor. G4NN had the highest number of true positive and true negative predictions, giving it the best combination of high sensitivity and high specificity of all predictors. The cG/cC score and RNAfold have similar levels of sensitivity and specificity, but had higher numbers of false positives than did G4NN. As is observable by the comparison of the G-tracts with the "+" sign predicted by RNAfold with those actually observed by *in-line* probing (boxed G-tracts in **Figure 34**), RNAfold often correctly predicted rG4 folding, but with wrong G-tracts. G4H was originally developed for DNA sequences, which could explain its lower sensitivity for RNA PG4 sequences. However, it showed high specificity, identifying only real rG4, just not all of them. Of the 26 candidates, only 7 were correctly predicted by all of the predictors (rG4: FZD10, FZD2, BMPR1A and MAPK3, dsRNA: BAG-5, SMAD2 and SMAD7). These well-predicted rG4 shared the characteristics of having short G-tracts that are separated by short loops of the same sizes (**Figure 34**). The well predicted dsRNA candidates presented longer loops and possessed a higher number of consecutive cytosines in order to compete with rG4 formation, and this was well detected by all of the predictors. Conversely, four candidates were wrongly predicted by all of the predictors (CASP9, BAD, BCL-2 and APPL1). These sequences presented larger loop sizes and many G-tracts of various lengths (**Figure 34**). These features still represent challenges that will need to be addressed for accurate prediction of rG4 folding motifs. The assessment of folding with a technique such as *in-line* probing permits the identification not only of the global rG4 or dsRNA structures, but also of which exact nucleotides are base-paired or not. Consequently, the experimental validation of folding remains essential to observe the limits of the actual prediction tools.

**Most rG4 possess features different than the canonical motif**

The *in-line* probing cleavage patterns permit the identification of an rG4 region in a given sequence with all of the possible combinations of four protected G-tracts and three loops with flexible nucleotides (**Figure 34**). However, this pattern represents the sum of all of the rG4 conformations in solution, and cannot identify which one is the most stable or dominant. **S1 Fig, panel B in Annexe 4** presents the *in-line* probing results of different G/A-mutant constructions for the MAPK3 sequence that can support the claim of two consecutive rG4

folding units like the beads-on-string model (Martadinata et Phan, 2014). Considering all candidates, the majority present more than four G-tracts. This indicates that multiple combinations of G-tracts can be adopted in order to fold into a single rG4. To avoid selecting a "preferred conformation", all possible G-tracts located immediately 5' or 3' of nucleotides with a $K^+/Li^+$ cleavage ratio higher than the threshold of 2 were boxed in the **Figure 34**. Each supplementary G-tract, adds numerous possibilities of different G-tract combinations. For example, the BAD candidate possesses seven G-tracts, giving rise to 35 possible combinations for rG4 formation.

The sizes of the G-tracts are an important feature of rG4, as the number of Gs comprising the G-tracts represents the number of stacked quartets of the resulting intramolecular quadruplex. The WNT set of positive rG4 candidates presents a lower number of G-tracts composed of 2Gs, and a higher number of G-tracts with 5Gs or more, in comparison with both the Apoptosis, the TGF-β and the PI3-K sets. As expected, the majority of the G-tracts are 3Gs in size as was defined in the initial sequence motif search. Interestingly, less than half of the positive rG4 candidates possessed four G-tracts of the same size. This means that the G-quartets of the structure are formed from G-tracts of various lengths, and that either not all of the Gs from the same tract are used simultaneously, or bulges might be present. Again, this result demonstrates why sequence motif searches and rG4 predictions based primarily on finding sequences with four identical lengths of G-tracts is not accurate.

The second most important feature that is required in order to define an rG4 is the loops linking the G-tracts. By broadly defining the rG4 loops as the nucleotides linking the protected G-tracts from the *in-line* probing cleavage pattern in the $K^+$ condition, one can observe that the rG4 candidates present loops that can vary greatly from the size of 1 to 7 nt defined in the canonical rG4 motif. The APPL1, BAD, BAG-1, CASP8AP2 and MAPK3 candidates *in-line* cleavage patterns allow for possible loops larger than 7 nts (**Figure 34**). Another distinctive feature of the loops is their nucleotide composition. No differences were observed in the ratios of A, U, C and G nucleotides in the loops. However, the presence of three consecutives Cs or Gs, in the loops, which could be respectively detrimental or beneficial for rG4 formation is highlighted in **Figure 34**.

These specific features, such as the sizes of the G-tracts and the sizes of the loops, might affect the stabilities of the different rG4s. For example, a stack of four quartets is more stable than a stack of three, and shorter loops are more stable than longer ones. Moreover, a combination of distinct features (G-tract numbers and sizes; loop numbers, sizes and composition) could also serve as motifs for recognition by *trans*-factors or helicases. The WRN DNA G4 helicase has been shown to bind G4s that are located in promoters. The helicase targets G4s that possess specific features that are different from those recognized by another DNA G4 helicase, specifically BLM (Tang *et al.*, 2016). Consequently, a similar recognition of specific features for rG4-helicase is a possibility. The development of chemical ligands with which to specifically target the rG4 structures is now a thriving field. The identification of specific features allows one to steer the design of a ligand such that it can target specific subsets of rG4 more precisely.

**rG4 folding affects the expression level of a luciferase reporter gene in colorectal cancer cells**

The *in vitro* experiments confirmed that rG4s are folded in the 5'UTR of many mRNAs involved in dysregulated colorectal pathway, suggesting a possible regulatory role for the structure. However, the *in cellulo* impact of the structure on the regulation of gene expression cannot be directly inferred from evidence of *in vitro* formation. In order to evaluate the potential role of the 5'UTR rG4s on mRNA regulation of expression, three selected rG4 candidates associated to different pathways (BAG-1 and CASP8AP2 for Apoptosis and MAPK3 for PI3-K) were selected for a gene reporter assay (Halder *et al.*, 2012). These candidates were selected in order to have at least one representative candidate for each pathway enriched with rG4 forming sequences. The TGF-β pathway was not further considered as "enriched" for rG4 structure because it contains only 2 positive rG4s out of 7 sequences. The candidates were the ones with the highest scores for each of the four *in silico* prediction tools and clear *in-line* probing and NMM fluorescence confirmations of rG4 formation. Importantly, these candidates have not been previously tested *in cellulo*. To ease the cloning procedures, the selected candidate also possessed short 5'UTR sequence. The three selected candidates were compared for their *in cellulo* effect on luciferase expression with candidates from the WNT (APC and FZD2) and the Apoptosis (BCL-2) pathways

already evaluated in previous work (Beaudoin et Perreault, 2010 ; Jodoin *et al.*, 2014 ; Shahid *et al.*, 2010, p. 2).

In order to perform the *in cellulo* assay, the entire 5'UTR containing the rG4 of the candidate was inserted upstream of a *Renilla Luciferase* (*Rluc*) reporter gene, and the resulting construct was then transfected into HEK293 cells. In parallel, the full length 5'UTR, bearing the same G/A-mutations confirmed to be negative for rG4 formation in the *in vitro* assays was also transfected in order to compare the effect of rG4 abolition on the expression level. A second plasmid coding for the *Firefly Luciferase* (*Fluc*) was co-transfected for normalization purposes. For all four candidates the *Rluc* normalized expression level was higher, with an approximately 2-fold increase for the G/A-mutants, in which rG4 formation was abolished, as compared to the WT (**S3 Fig in Annexe 4**). The formation of an rG4 in these four candidates represses expression, as was observed previously for many other rG4s located in the 5'UTR (Halder *et al.*, 2012; Song *et al.*, 2016).

The *in-line* probing results suggests that the MAPK3 candidate formed two adjacent rG4s. In order to confirm this observation, different G/A-mutants where designed so as to abolish either the first possible rG4, the second, or both of them. As seen in **S3 Fig, panel D in Annexe 4**, the two G/A-mutants impairing a single rG4 have similar *Rluc* normalized expression levels that are almost 3-fold higher than that of the WT sequence, but the double rG4 mutant exhibits a 6-fold increase in the expression level of the *Rluc* reporter gene over that of the WT. This indicates that there is an additive, repressive effect of the two rG4. Even with the double rG4 mutant, in which 9 Gs were mutated to As, not all possible rG4 formations were abolished. The *in-line* probing (**S1 Fig, panel B, Annexe 4**) and NMM fluorescence assay (**S2 Fig, Annexe 4**) results of MAPK3 demonstrated that the double G/A-mutant can still adopt an rG4, albeit possibly a less stable one possessing only 2 stacked quartets and loops sizes of 3 and 5 nts. Thus, the 6-fold increase in the expression level of the double rG4 mutant over the WT observed *in cellulo* might actually be higher if all potential rG4 were eliminated.

Gene-reporter assays using HEK293 cells permit comparisons with previous studies in which rG4 located in 5'UTRs were evaluated for their repressive effects on expression levels *in cellulo*. The fold changes of the normalized luciferase expression of the mutant over WT (2.40-fold for BAG-1 and 2.07-fold for CASP8AP2) are in the same range as the ones

observed for the APC, BCL-2 and FZD2 rG4s in earlier studies (1.74-fold for APC(Jodoin *et al.*, 2014), 2.30 -fold for BCL-2 (Shahid *et al.*, 2010, p. 2) and 2.50-fold for FZD2 (Beaudoin et Perreault, 2010)). However, because the mRNA 5'UTR rG4 candidates selected here were associated with colorectal cancer dysregulated pathways, the gene-reporter assays were also performed using three representative colorectal cell lines: HCT116, HT29 and DLD-1 for the APC, BAG-1 and CASP8AP2 candidates (**S4 Fig. in Annexe 4**). The MAPK3 candidate was not further tested in colorectal cell lines, as no complete rG4 negative control was possible without highly mutating the short 5'UTR sequence with supplementary G-to-A mutations.

In general, the results replicate what was observed with the HEK293 cells. In the colorectal cell lines the normalized expression of the *Rluc* reporter gene was higher when the rG4 was mutated for all three of the candidates tested. Despite slight differences in the expression levels, the mutant over WT fold-changes were very similar for a specific candidate between the three cells lines. The APC G/A-mutant fold-change was ~2 (2.09-fold in HCT116, 2.11-fold in HT29 and 2.65-fold in DLD-1) and the BAG-1 fold-change was ~3 (3.56-fold in HCT116, 3.11-fold in HT29 and 3.92-fold in DLD-1). The difference was not statistically significant for the CASP8AP2 candidate, which showed an almost 2-fold increase in the expression level for the mutant in all three cell lines (1.87-fold in HCT116, 1.95-fold in HT29 and 2.01-fold in DLD-1). In brief, rG4 folding in the 5'UTR of these mRNAs associated with colorectal cancer dysregulated pathways represses the expression level of the luciferase reporter gene in the relevant colorectal cancer cells models.

The *in cellulo* assays performed cannot decipher at which level, transcriptional, post-transcriptional or both, the regulation occurred. However, based on actual knowledge of rG4 regulation in 5'UTRs (Bugaut et Balasubramanian, 2012), translational repression seems to be the probable mechanism. The range of repression levels between the different rG4 candidates, and between the different cell lines, observed here and in other studies is quite narrow, generally being a 2-to-3 fold-change between the WT rG4 and the mutated sequence. Nevertheless, the different levels of repression observed between the candidates might be explained by how the different features of the rG4s themselves (i.e. the G-tracts and loops) affect their stabilities, and also by the differences in their full 5'UTR contexts (i.e. the position where the rG4 is located in the 5'UTR and exactly what are the adjacent sequences).

Bhattacharyya *et al.*(Bhattacharyya *et al.*, 2017) analyzed the role of the context by interchanging two rG4s that exhibited opposite effects on expression (one was an enhancer, the other a repressor). After the switch, the rG4 in the new position mimics the effect on expression level of the original rG4. It showed that the context is also responsible for the rG4 mechanism of regulation. For a particular rG4 candidate in its natural 5'UTR context, the differences in the levels of repression between the alternate cell lines might be caused by variations in the *trans*-factors that are expressed in those cell lines and that take part in the rG4-mediated regulation of mRNA expression. In order to continue with the hypothesis that rG4s located in the 5'UTR are part of a global regulation mechanism, it would be interesting to compare the characteristics of the 5'UTR in which they are found for other similarities (position of the rG4, upstream ORFs, IRES, translation regulatory sequences, protein-binding motifs, etc.) that could also be specific to each pathway.

## CONCLUSION

The enrichment of rG4 prone sequences in the 5'UTRs of mRNAs, and their known repressive effects on translation, point towards a possible role for these structures in the global regulation of mRNAs involved in common biological pathways. This study showed that mRNAs bearing a consensus PG4 sequence in their 5'UTRs are enriched in some of the KEGG annotated pathways, for example in the general colorectal cancer pathway. *In vitro* evaluation of the folding of 26 selected PG4 candidates associated with well-defined colorectal cancer dysregulated pathways confirmed rG4 folding for 15 of them and *in cellulo* reporter assays using colorectal cell lines demonstrated their effect on mRNA expression level for 3 of them.

This study adds new, experimentally confirmed sequences to the list of rG4s located in the 5'UTR that could affect gene expression. It demonstrates that rG4 prediction based solely on a sequence motifs search is insufficient. The available *in silico* prediction tools, such as G4NN which was the best one for the set of candidates examined here, can improve the selection of rG4 prone sequences, but cannot yet correctly predict which sequences will fold, nor which exact nucleotides of the sequence are involved in the structure.

*In-line* probing of rG4 sequences permits identification of the nucleotides involved in quadruplex formation, and thus comparison of their features (e.g. G-tract numbers, sizes,

loops sizes and compositions). However, further studies are needed in order to uncover specific rG4 features shared by mRNAs involved in a common biological pathway, and to better understand the role of both helicases and rG4 binding proteins in the recognition mechanism of subsets of distinct rG4s.

## SUPPLEMENTARY DATA

### Annexe 4 :

**Supplementary tables S1–S5**

**S1 Table.** Sequences, positions in the 5'UTR and lengths of all candidates and their respective full-length 5'UTRs.

**S2 Table**. Comparison of the prediction methods.

**S3 Table.** UTRref, RefSeq and Gene-ontology Identification numbers of all candidates.

**S4 Table.** Oligonucleotide sequences used for PCR-filling prior to in vitro transcription.

**S5 Table**. Oligonucleotide sequences used for PCR filling prior to cloning.

**Supplementary figures S1–S4**

**S1 Fig**. *In-line* probing gels and $K^+/Li^+$ ratio quantification of the candidates.

**S2 Fig.** NMM assay of all candidates.

**S3 Fig**. *In cellulo* luciferase assay in HEK293 cells.

**S4 Fig.** *In cellulo* luciferase assay in colorectal cancer cell lines.

## ACKNOWLEDGMENTS

We thank Jean-Denis Beaudoin for the initial PG4 database, Jean-Michel Garant for the use of G4RNAscreener, Cameron Levins for technical assistance, Nathalie Rivard and Étienne Lemieux both for access to their colorectal cancer cell lines and for their expertise and Martin Bisaillon for critical reading of the manuscript.

# ARTICLE 5 – G-QUADRUPLEX LOCATED IN THE 5'UTR OF THE BAG-1 MRNA AFFECTS BOTH ITS CAP-DEPENDENT AND CAP-INDEPENDENT TRANSLATION THROUGH GLOBAL SECONDARY STRUCTURE MAINTENANCE

**Auteurs de l'article**: Jodoin, Rachel, Carrier, Julie, Rivard, Nathalie, Bisaillon, Martin and Perreault, Jean-Pierre

**Statut de l'article** : Soumis à Nucleic Acids Research, 27 novembre 2018

**Avant-propos** : Rachel Jodoin a effectué le design expérimental, a réalisé les expériences et les analyses. Julie Carrier a fourni l'accès à la biobanque de tissus pairés. Nathalie Rivard a fourni l'accès aux lignées cellulaires colorectales et conseillé au design des expériences initiales de régulation post-transcriptionnelle de BAG-1 en tissus et lignées colorectales. Martin Bisaillon a conseillé le design des expériences de synthèse de la coiffe, d'ARNm et de gènes rapporteurs. L'article a été rédigé par Rachel Jodoin et Jean-Pierre Perreault, et révisé par Martin Bisaillon et Nathalie Rivard.

**Résumé**

La protéine anti-apoptotique BAG-1 est connue pour être surexprimée dans les tumeurs colorectales. Son ARNm encode pour trois isoformes protéiques principaux résultant de l'initiation de la traduction à quatre codons de départ alternatifs, présents dans le même cadre de lecture et situés dans des contextes d'initiation sous-optimaux. La région 5'UTR contient aussi un site interne d'entrée du ribosome (*internal ribosome entry site*, IRES) qui régule la traduction coiffe-indépendante de l'isoforme le plus court. La formation *in vitro* d'un G-quadruplex d'ARN avait déjà été confirmée à l'extrémité 5' du 5'UTR de BAG-1, en amont de tous les éléments de régulation en *cis* déjà caractérisés. Puisque les rG4 situés en 5'UTR agissent principalement comme répresseurs de la traduction et puisque certains rG4 particuliers ont déjà été identifiés comme étant impliqués dans la formation et la régulation

de structure secondaire IRES, les rôles du rG4 de BAG-1 sur les traductions de types coiffe-dépendante, et indépendante, ainsi que son interaction avec les multiples codons d'initiation ont été examinés. La régulation de la traduction de l'ARNm BAG-1 par la présence d'un cadre de lecture ouvert répressif en amont (*upstream ORF,* uORF) est décrite ici pour la première fois. En utilisant une combinaison de gènes rapporteurs et d'analyses de cartographie *SHAPE* du 5'UTR complet, le rG4 a démontré des effets opposés sur les deux types de traductions. Ce résultat s'explique par l'impact du repliement rG4 sur la structure secondaire globale.

**Abstract**

The anti-apoptotic BAG-1 protein is known to be overexpressed in colorectal tumors. Its mRNA encodes for three protein isoforms resulting from alternative translation initiation that occur at four in-frame start codons located in suboptimal translation initiation contexts. The 5'UTR also contains an internal ribosome entry site (IRES) that regulates the cap-independent translation of the short isoform. An RNA G-quadruplex (rG4) was previously confirmed to fold *in vitro* at the 5'end of the BAG-1 5'UTR, upstream of all known *cis*-regulatory elements. As rG4s located in the 5'UTR are known to act mostly as translational repressors, and because individual rG4s have been shown to be involved in IRES secondary structure formation and regulation, the role of the BAG-1 rG4 on both cap-dependent and independent translation of the BAG-1 mRNA, and its interplay with the multiple start codons, was investigated. The regulation of the translation of the BAG-1 mRNA by a repressive upstream ORF is described for the first time. Using a combination of reporter genes and whole 5'UTR structural probing by SHAPE, the rG4 was shown to exhibit opposite effects on the two types of translation, a result that is explained by the impact of its folding on the global secondary structure.

## INTRODUCTION

The BAG-1 protein (Bcl2-associated athanogene 1) was initially identified as an interactor of the anti-apoptotic protein BCL-2 (Takayama *et al.*, 1995), and is known to be an inhibitor of the intrinsic apoptotic pathway (Tang, 2002; Wang *et al.*, 1996). BAG-1 was further characterized as being a multifunctional protein. Among its functions, BAG-1 acts as a nucleotide exchange factor that modulates the activity of the chaperones Hsp70/Hsc70 (Alberti *et al.*, 2003; Lüders *et al.*, 2000). BAG-1 is also known to interact with a diverse array of partners including the retinoblastoma protein (pRb) (Clemo *et al.*, 2005); the oncogenic kinase Raf-1 (Song *et al.*, 2001); several nuclear hormones and growth receptors such as the platelet-derived growth factor receptor (PDGF-R), the hepatocyte growth factor receptors (HGF-R), the androgen receptor, and, NFκB (Clemo *et al.*, 2008; Townsend *et al.*, 2003; Wood *et al.*, 2009; Zeiner et Gehring, 1995), to name a few. Overall, depending on the interactions with the various partners, BAG-1 integrates signals from multiple pathways resulting in phenotypes of cell proliferation, growth, survival, transcription regulation and protein modifications (Townsend *et al.*, 2005).

The BAG-1 protein is expressed in three main isoforms: the long, BAG-1L (50 kDa), the medium, BAG-1M (46 kDa), and, the short, BAG-1S (36 kDa). There is also a fourth isoform, less abundant, BAG-1 p29 (29 kDa). All BAG-1 protein isoforms are translated from the same mRNA transcript using a mechanism of alternative translation initiation called leaky scanning at the four alternative in-frame start codons present in the 501 nucleotides (nts) long 5'-untranslated region (5'UTR) of the mRNA (Packham *et al.*, 1997; Yang *et al.*, 1998) (**Figure 35A**). The resulting products of protein synthesis are thus protein isoforms that differ in the length of their amino (N)-terminal extensions (**Figure 35B**). All isoforms possess both the ubiquitin binding ligand (UBL) and the BAG domains at the C-terminal end, domains that are essential for protein-protein interaction with most of the known partners. The isoforms differ in the number of acidic repeats found in the N-terminal region. The BAG-1L isoform is the only one that possesses a nuclear localisation signal (NLS) at its N-terminal end that triggers its localization in the nucleus. BAG-1M shuttles between the nucleus and the cytoplasm, and BAG-1S, the most abundant isoform, is cytoplasmic.

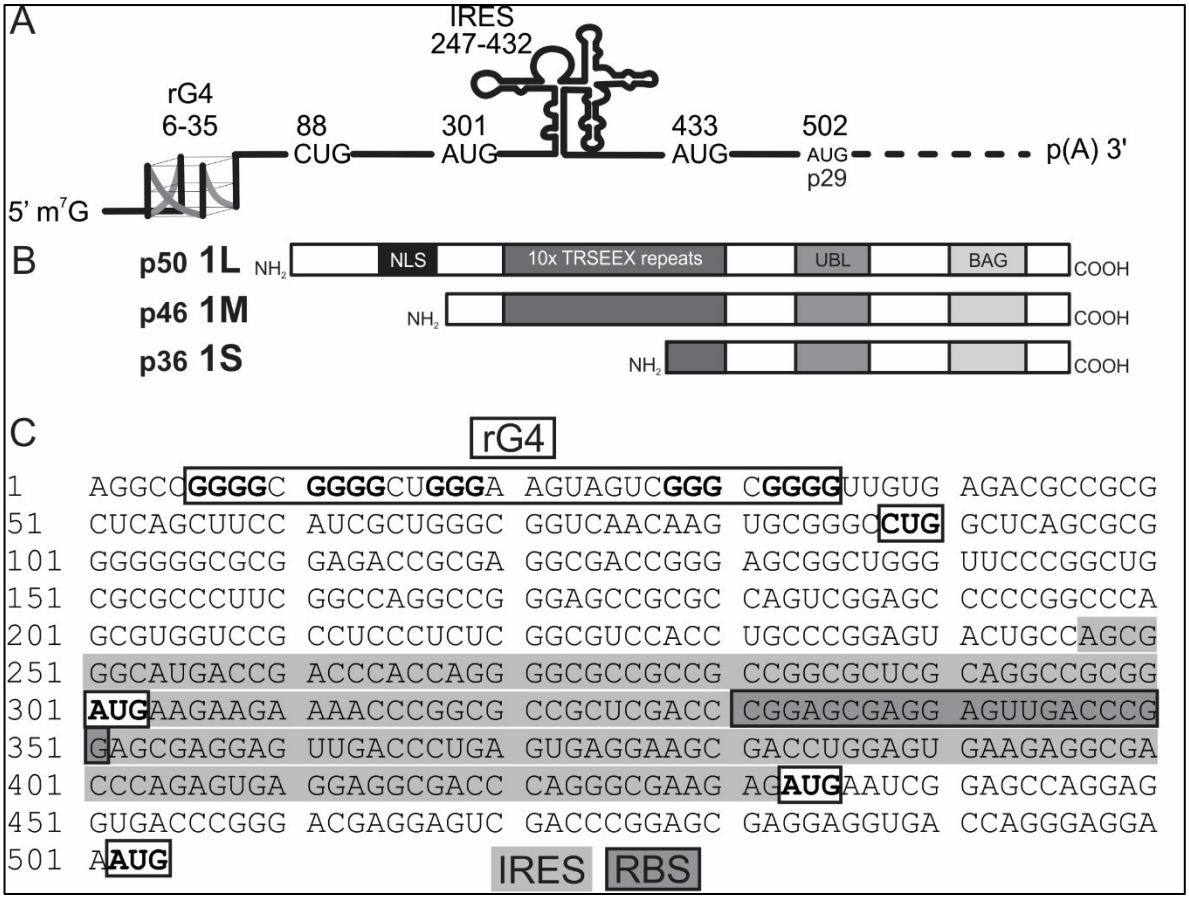



**Figure 35** – Scheme of the BAG 1 5'UTR organization.

(A)The BAG-1 mRNA presents many features in its 5'UTR: an rG4 secondary structure located at its 5' end; four in-frame start codons with the first being a non-canonical CUG; and, an IRES secondary structure. (B) Translational initiation at the three principal alternative start codons results in the production of three protein isoforms (1L, Long; 1M, Medium; and, 1S, Short) diverging from each other by the size of their N-terminal extension. (C) Nucleotide sequence of the complete BAG-1 5'UTR. The rG4 region, the start codons, the IRES region and the ribosome binding sites are highlighted.

The different isoforms are known to be overexpressed in many different cancers (Sharp *et al.*, 2004), including colorectal cancer (CRC) (Clemo *et al.*, 2008; Skeen *et al.*, 2013; Southern *et al.*, 2012). The protein's expression is the highest in the late stages of colorectal tumorigenesis (Clemo *et al.*, 2008) and the overexpression of the long isoform BAG-1L is associated with a poorer prognosis (Barnes *et al.*, 2005; Kikuchi *et al.*, 2002). Notably, the treatment of CRC cells with BAG-1 directed small silencing RNA (siRNA) induces apoptosis (Huang *et al.*, 2016; Xiong *et al.*, 2003). Thus, BAG-1 is considered as being a possible

therapeutic target in CRC (Collard *et al.*, 2012), as well as in many other cancer types (Aveic *et al.*, 2015; Cato *et al.*, 2017; Papadakis *et al.*, 2016).

The BAG-1 5'UTR contains several regulatory features at both the RNA sequence and the secondary structure levels (**Figure 35C**). First, the four alternative in-frame start codons are all located in suboptimal Kozak contexts for translation initiation. Second, the translation of the longest isoform, BAG-1L, is initiated at a non-canonical CUG start codon. Third, the translation of the most abundant isoform, BAG-1S, is regulated by an internal ribosome entry site (IRES)(Coldwell *et al.*, 2001). The nucleotides located at positions 247 to 432 of the 5'UTR adopt a defined secondary structure element that recruits the IRES *trans*-acting factors (ITAF) PTB-1 and PCBP1. This allows the opening of a ribosome binding site (RBS) window and the recruitment of the 40S ribosomal subunit inside the 5'UTR, thereby initiating translation in a manner independent of 5' cap-recognition (Pickering *et al.*, 2004). This regulation of the BAG-1S isoform happens under stress-related conditions such as heat-shock and chemotoxic stresses (Coldwell *et al.*, 2001; Dobbyn *et al.*, 2008) where the canonical cap-dependent translation is repressed. Finally, our group previously identified and probed *in vitro* a G-quadruplex secondary structure at the 5' extremity of the 5'UTR of the BAG-1 mRNA, specifically at the positions 6 to 35 (Jodoin *et al.*, 2014) (**Figure 35A and C**), that repressed the expression of a luciferase reporter gene in three CRC cell lines (Jodoin et Perreault, 2018).

G-quadruplexes (G4) are very stable non-canonical secondary structures formed by G-rich DNA or RNA nucleotide sequences. In a sequence presenting a minimum of four tracts of two or more continuous guanines, each guanine of the tract can form base-pairs with the ones from the next tracts through Hoogsteen hydrogen bonds, resulting in a co-planar array called the G-quartet. The stacking of two or more G-quartets forms a G4. In intramolecular G4, the G-quartets are linked to each other via three stretches of any nucleotides that form the loops. The G4s are stabilized by the presence of a monovalent cation, mainly potassium, the most abundant cation in the cell.

RNA G4 (rG4) are highly abundant (Kwok *et al.*, 2016a) and are folded *in cellulo* (Biffi *et al.*, 2014a). They are involved in many post-transcriptional regulation mechanisms such as alternative splicing, polyadenylation and mRNA localization (Millevoi *et al.*, 2012). They are specifically bound by RNA-binding proteins and helicases in order to regulate their

formation (Fay *et al.*, 2017), and are involved in various diseases (Cammas et Millevoi, 2017). The rG4 located in the 5'UTR have principally been identified as translational repressors (Beaudoin et Perreault, 2010). The proposed mechanism is by sterically blocking both the translation initiation and the ribosomal scanning because of their high stability (Bugaut et Balasubramanian, 2012). However, in contrast to the majority of rG4s, the rG4s located in the 5'UTRs of the VEGF and the FGF-2 transcripts were identified as being parts of IRES secondary structures that are essential for the cap-independent translation of these mRNAs (Bonnal *et al.*, 2003; Morris *et al.*, 2010). Despite their high 5'UTR abundance, and their known role in translational regulation, the possible interplay of rG4s with other *cis* translational regulation motifs located in 5'UTRs such as alternative start codons, non-canonical start codons and IRES, is currently overlooked.

The BAG-1 mRNA transcript's translation is regulated by both non-canonical cap-dependent and cap-independent translation mechanisms. The impacts of a 5'UTR rG4 on the translational regulation of a transcript where both types of regulation are simultaneously present has never been deciphered. The rG4 region is located upstream of all known regulatory elements in the BAG-1 5'UTR. The effect of the rG4 folding on the secondary structure and on the post-transcriptional regulation of the 5'UTR in its full context was thus investigated in detail using various luciferase reporter constructs and complete 5'UTR structural probing.

## MATERIAL AND METHODS

### Paired colorectal tumor tissue samples

Total protein lysates in RIPA buffer (25 mM Tris-HCl pH 7.6, 150 mM NaCl, 1% NP-40, 1% sodium deoxycholate, 0.1% SDS) and complementary DNAs (cDNA) resulting from reverse-transcription (RT) of the total RNA extracted from 50 specimens of paired tumoral and healthy colorectal tissues were obtained from a biobank previously described (Bian *et al.*, 2016). The healthy tissue consists of the margin located at least 10 cm away from the tumor. The tissues were obtained from patients, who did not received neoadjuvant therapy, undergoing surgical resection. The tissues were processed, classified and graded as previously described (Bian *et al.*, 2016). The clinicopathological parameters of the patients and tumors are described in **Supplementary Table S1 in Annexe 5**. The protocol was

approved by the Institutional Human Subject Review Board of the Centre Hospitalier Universitaire de Sherbrooke and the patients' written, informed consents were obtained.

The BAG-1 mRNA levels were determined by qPCR in human advanced adenomas and adenocarcinomas and were compared to the paired adjacent healthy tissue for 46 samples (Adenoma n=8; Stage 1 n=8; Stage 2 n=10; Stage 3 n=10; and, Stage 4 n=10). The BAG-1 protein isoform levels were determined by Western blot analysis for 38 pairs of samples (Adenoma n=7; Stage 1 n=7; Stage 2 n=8; Stage 3 n=8; and, Stage 4 n=8). Only a small number of tissue pairs were not in common between both analyses (Adenoma n=3; Stage 1 n=1; Stage 2 n=2; Stage 3 n=2; and, Stage 4 n=2).

**Cell culture**

The HCT116 colorectal cancer cell line (ATCC, CCL-247) was cultivated in McCoy's 5A medium supplemented with 10 % foetal bovine serum (FBS) in a 37°C incubator with a 5 % $CO_2$ atmosphere. All cell culture reagents were obtained from Multicell, Wisent.

**Treatment of cells with G4-specific chemical ligands**

HCT116 cells were seeded at 650 000 cells/well in 6-well plates, 24 h prior to treatment. Along with 1µL/well of lipofectamine 2000 (ThermoFisher), the ligands were then added to the media at a concentration of 2 µM cPDS (Carboxypyridostatin trifluoroacetate salt, Sigma-Aldrich, working solution 1 mM in water), 20 µM Phen-DC3 (Polysciences Inc., working solution 2 mM in DMSO) and 2 µM TmPyP4 (meso-5,10,15,20-Tetrakis-(N-methyl-4-pyridyl)porphine, Calbiochem, working solution 1 mM in water) and the cells incubated during 24 h at which point they were compared to vehicle-only treated cells. All treatments were performed in triplicate (for cPDS) or in duplicate (Phen-DC3 and TmPyP4), and were repeated at two different days ($n$=2). Cells from each well were harvested in 1 mL of ice-cold PBS using a cell scraper. The cell volumes equivalent to 1/5 and 4/5 of a well of a 6-well plate were kept for total RNA and total protein extractions, respectively. Centrifugation at 1 000 RPM for 10 min was performed to isolate the cell pellets, which were stored at -80°C until the lysis and further RNA and protein extractions were performed.

**Design and cloning of the gene reporter constructs**

*PsiCHECK-2 luciferase reporter*

The complete WT, rG4mut and the 1S start codon mutated sequences of the BAG-1 5'UTR with flanking NheI restriction sites were chemically synthesized and ordered from Biomatik. After NheI digestion, the 5'UTR was ligated to the psiCHECK-2 dual-luciferase reporter plasmid (Promega) upstream of the Rluc coding sequence (CDS). To ensure that all start codons were in-frame with the Rluc CDS, two extra nucleotides were added by primer directed mutagenesis. The 1L, 1M and AUG-254 start codons were mutated using primer directed mutagenesis. All sequences were verified by DNA sequencing.

*pRL-HL bicistronic luciferase reporter*

The pRL-HL bicistronic luciferase reporter plasmid consists of the Rluc reporter gene, expressed via cap-dependent translation, followed by the NotI restriction site, all located upstream of the HCV IRES sequence that controls the Fluc expression in cap-independent fashion. The bicistronic plasmid was modified by directed mutagenesis to insert a HpaI restriction site at the 3' end of the HCV IRES. The removal of the HCV IRES sequence was performed by digesting the vector with the NotI and HpaI restriction enzymes. Both the BAG-1 complete 5'UTR WT or rG4 mutant were amplified from the psiCHECK-2 constructions using primers inserting both the NotI and HpaI restriction enzyme sites, and were then digested and ligated in between the two luciferase reporter genes to create either the pRL-BAG1wt or the G4mut-HL bicistronic vectors. Other mutations in the BAG-1 IRES sequence, stem3mutA and stem3mutB, were generated using primer directed mutagenesis and were verified by DNA sequencing.

The complete list of sequences used in this study, as well as the list of primers, are available in **Supplementary Tables S3 and S4 in Annexe 5**.

**Transfections and luciferase assays**

*Transfection of the monocistronic psiCHECK-2 luciferase reporter construct*

Twenty-four hours prior to transfection, HCT116 cells were seeded at 650 000 cells/well in 6-well plate. The cells were transfected using 125 ng/well of the psiCHECK-2 construction along with 2375 ng/well of the carrier plasmid PUC19 using 2.5 µL/well of lipofectamine

2000 (ThermoFisher*)* in serum-free media. The serum was added 4 h after transfection. Twenty-four hours later, the cells were harvested on ice using 1 mL of PBS 1X and a cell scraper. The cell lysate was divided in 3 parts: 1/5 (200 µL) for qPCR, 1/5 (200 µL) for the luciferase assay and 3/5 (600 µL) for Western blot.

*Transfection of the bicistronic pRL-HL luciferase reporter construct*

Twenty-four hours prior to transfection, HCT116 cells were seeded at either 300 000 cell/well in 12-well plates, or at 100 000 cells/well in a 24-well plates. The cells in the 12-well plates were transfected using 1000 ng/well of the bicistronic constructions (500 ng/well for the 24-well plates) using 2 µL/well of lipofectamine 2000 (1µL/well for the 24-well plates) in serum-free media. The serum was added 4 h after transfection. The cells were harvested 24 h later. For the experiments performed in the 12-well plates, half of the cells (500 µL) were used for qPCR and half for the luciferase assay. In the case of 24-well plates, all of the cells were recovered and lysed to perform the luciferase assay.

*Luciferase assay*

The DualGlo luciferase assay kit from Promega was used according to the manufacturer's protocol. Briefly, the cells were lysed in the corresponding cell volume amount of the kit's 1X passive lysis buffer. A volume of 5 to 10 µL of the cell lysate was used, and 100 µL of each of the luciferase substrates was added sequentially. Readings of 5 sec integration times were performed using a Glomax 20/20 luminometer.

**Western blot**

*Endogenous BAG-1 isoforms in colorectal tumor and paired margin*

Proteins (20 µg) obtained from the total protein lysate of the tissue samples obtained from the biobank were separated on a 10% SDS-PAGE gel and then was transferred to a polyvinylidene difluoride (PVDF) membrane. The membrane was blocked 30 min at room temperature in Tris buffered saline (TBS) with 2.5 % (w/v) nonfat dry milk, then incubated overnight (O/N) at 4°C with the primary antibody mouse mAb anti-BAG-1 (CC9E8, Santa Cruz Biotechnologies), diluted 1:100 in phosphate buffered saline (PBS) with 2.5 % (w/v) nonfat dry milk (PBS-milk 2.5 %). After three washes in PBS with 0.1 % Tween-20 (PBS-T), the membrane was incubated for 1 h at room temperature with the secondary antibody

anti-mouse IgG (H+L) Alkaline phosphatase-conjugated antibody (Promega), at a dilution of 1:7 500 in PBS-milk 2 %. After two washes for 10 min with PBS-T, and one with PBS only, the membrane was developed using 1 mL of alkaline phosphatase substrate (CDP-Star (Applied Biosystems) diluted to 1X in 100 mM Tris pH 9.5 and 100 mM NaCl buffer). The membrane was rinsed with PBS-T and then exposed to an x-ray film for diverse exposure times. The loading control ERK2 was obtained after 2 h of incubation at 37°C of the membrane with the rabbit anti-ERK2 antibody (C14, Santa Cruz Biotechnologies) diluted 1:5 000 in PBS-milk 5 %. After washes with PBS-T, the membrane was incubated for 1 h at room temperature with the secondary anti-rabbit L-HRP antibody (NA934, GE Healhcare) diluted 1:5 000 in PBS-milk 5 %. Revelation was performed using the Western lightning plus-ECL enhanced chemiluminescence substrate (PerkinElmer), and was detected using the ImageQuant LAS4000 (GE Healthcare). Quantification of band densities was obtained using the ImageJ software.

*Endogenous BAG-1 in HCT116 cells treated with ligands*

The cell lysis for total protein extraction was performed by the addition of 100 µL of Laemmli buffer 1.5X (3.75% SDS, 15% Glycerol, 150 mM Tris-HCl pH 6.8) per cell pellet that corresponded to 4/5 of a well of a 6-well plate of treated cells. The samples were boiled for 5 min at 90°C, and were then sonicated twice for 2 sec at 16 % amplitude. The samples were then centrifuged for 1 min at 13 000 RPM, and the protein concentration in the supernatant was evaluated using the BCA assay (Pierce) according to the manufacturer's protocol. Protein lysates (30 µg) were loaded on a 10% SDS-PAGE gel and the Western blot against the endogenous BAG-1 protein was performed as described above. After membrane stripping with two washes for 10 min each with NaOH 0.5 N and one wash for 10 min in PBS, the membrane was blocked for 20 min in PBS-milk 2.5%. The membrane was then incubated O/N at 4°C with the loading control antibody anti-β-actin mouse mAb (A5441, Sigma) diluted 1:1 000 in PBS-milk 2.5%. After three washes for 10 min each with PBS-T, the membrane was incubated for 1 h at room temperature with the secondary anti-rabbit L-HRP antibody (NA934, GE Healhcare) diluted 1:5 000 in PBS-milk 5 %. Revelation was performed using the Western lightning plus-ECL enhanced chemiluminescence substrate (PerkinElmer), and was detected using the ImageQuant LAS4000 (GE Healthcare). Quantification of band densities was performed using the ImageJ software. The BAG-1

protein isoforms abundance levels were normalized over the abundance level of the β-actin loading control.

*N-terminal extended Rluc isoforms in transfected HCT116 cells*

Using a cell volume equivalent to 3/5 of a 6-well plate of transfected cells, the lysis was performed with 70 µL of Laemmli buffer 1.5X. Protein lysates (30 µg) were migrated on a 10 %, SDS-PAGE gel. The gel was then transferred to a nitrocellulose membrane. The membrane was blocked in PBS-Milk 4 % for 15 min at room temperature, then was incubated O/N in PBS-Milk 4 % with the primary antibodies rabbit polyclonal antibody Anti-Renilla Luciferase (PM047, MBL) diluted 1:1 000 and anti-β-actin mouse mAb (A5441 ,Sigma) diluted 1:2 000. The membrane was washed three times for 10 min with PBS-Tween 0.1 %, and then was incubated for 1 h at room temperature with the secondary antibodies Alexa fluor-680 Goat anti-Mouse IgG (A21057, Life technologies) and IRDye 800CW Donkey anti-Rabbit IgG (#926-32213, Mandel), both diluted 1:10 000 in PBS-Milk 4 %. After three washed for 10 min with PBS-T, the membrane was revealed using the Odyssey imaging system (LI-COR Biosciences). Quantification of band densities was obtained using the ImageJ software.

**Total RNA extraction of HCT116 cells and DNase treatment**

The total RNA extraction was performed using a cell volume corresponding to 1/5 of a 6-well plate, or 1/2 of a 12-well plate, depending on the experiment described in the previous sections. The cells were homogenized with 250 µL of QIAazol (QIAGEN). Then, the RNA extraction was performed by adding 50 µL of chloroform, incubating at room temperature for 2 min and centrifuging at 13 000 RPM for 15 min. The aqueous phase was then transferred into a new tube and the RNA was precipitated by the addition of 125 µL of isopropanol. After a 5 min incubation at room temperature and a centrifugation at 13 000 RPM at 4°C for 20 min, the resulting RNA pellet was washed with 225 µL of 70 % ethanol and centrifuged again at 13 000 RPM at 4°C for 10 min. The resulting pellet was air dried and dissolved in 30 µL of $H_2O$.

RNA samples were treated with DNase prior to RT-PCR. Briefly, 1 µg of total RNA was incubated in a final volume of 10 µL with 1 µL of 10X DNAse reaction buffer and 1 unit of RQ1 RNAse-free DNAse (both from Promega) for 30 min at 37°C. After incubation, 90 µL

of $H_2O$ were added and the RNA was recovered by phenol-chloroform extraction followed by ethanol precipitation. RNA pellet was dissolved in 5 µL $H_2O$ (resulting in a concentration of approximately 200 ng/µL) prior to be sent to the RNomics Platform of the Université de Sherbrooke to perform RNA quality control, reverse transcription and qPCR reactions.

**RNA Quality Control, Reverse Transcription and qPCR**

All of these steps were performed by the RNomics Platform of the Université de Sherbrooke. RNA integrity was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies). Reverse transcription (RT) was performed on 1.1 µg total RNA with final concentration of 10 units of Transcriptor reverse transcriptase, 60 µM of random hexamer, 1 mM each dNTP (all from Roche Diagnostics), and 10 units of RNAseOUT (*Invitrogen)* according to Roche Diagnostics' protocol in a total volume of 10 µL. All forward and reverse primers were individually dissolved to 20-100 µM stock solution in 10 mM Tris, 0.1 mM EDTA buffer (Integrated DNA technologies, IDT) and diluted as a primer pair to 1 µM in RNase DNase-free water (IDT). Quantitative PCR (qPCR) reactions were performed in a 10 µL volume in 384-well plates on a CFX-384 thermocycler (BioRad) with 5 µL of 2X iTaq Universal SYBR Green Supermix (BioRad), 10 ng (3 µL) cDNA and 200 nM final (2 µL) primer pair solutions. The following cycling conditions were used: 3 min at 95°C ; 50 cycles of: 15 sec at 95°C, 30 sec at 60°C, 30 sec at 72°C. The relative expression levels were calculated using the qBASE framework (Hellemans *et al.*, 2007) and the housekeeping genes YWHAZ, MRPL19, PUM1 and SDHA for human cDNA. Primer design and validation was evaluated as described elsewhere (Brosseau *et al.*, 2010). In every qPCR run, a no-template control was performed for each primer pair and these were consistently negative. All primer sequences are available in **Supplementary Table S3 in Annexe 5**.

**mRNA mono- and bi-cistronic luciferase reporter assays**

*Preparation of mRNA transcripts*

First, the pRL-intercistronBAG1wt or G4mut-HL plasmids were created starting from the pRL-BAG-1-HL reporter plasmids using primer directed mutagenesis. A 74 nts long intercistron region was added between the RLuc CDS and the complete 5'UTR sequence of BAG-1. Its function was to extend the 3'extremity after the Rluc CDS in order to augment the stability of the resulting Rluc monocistronic mRNA construct. The DNA templates used

for *in vitro* transcription to create both the mono- and bicistronic mRNA constructs (capped and poly-adenylated) for transfection were created by the amplification of either the pRL-intercistronBAG1wt or the G4mut-HL plasmid using different sets of primers. The primers were designed so as to add the T7 promoter in 5' and a 60 nts long poly-A tail in 3'. The primers used are listed in **Supplementary Table S3 in Annexe 5,** and the complete mRNA sequences of each construction are listed in **Supplementary Table S5 in Annexe 5**. After PCR amplification, the DNA templates were digested with the DpnI restriction enzyme to remove any remaining plasmid nucleotides, and were then purified using the PCR purification kit (Biobasic Canada inc.) according to the manufacturer's protocol.

*In vitro* transcription and capping of the mRNA with either the $m^7$G-cap or the analog A-cap was performed using the mMessage mMachine Kit (Ambion) according to the manufacturer's instructions. The only alteration to the protocol was to generate the mRNA constructions capped with the A-cap analog. The 2XNTP/CAP solution from the kit was replaced by a G(5')ppp(5')A RNA cap structure analog (NEB) to obtain the 2X NTP/Analog solution with final concentrations of 12 mM A-cap analog, 15 mM each of rATP, rCTP and rUTP, and 3 mM rGTP. After transcription, DNAse treatment and lithium chloride precipitation (following the manufacturer's protocol), samples of the mRNA constructions were verified on denaturing agarose gels for their integrity.

*mRNA transfection and luciferase assay*

HCT116 cells were seeded at 160 000 cells/well in 24-wells plates 24 h prior to transfection. A total amount of 500 ng of mRNA constructions (either 250 ng of the Fluc monocistronic constructions co-transfected with 250 ng of the RLuc monocistronic control, or 500 ng of a bicistronic construction) were transfected using 1 µL/well of lipofectamine 2000 in serum-free media. The cells were harvested 4 h after transfection. Half of the cell volume was used for the DualGlo luciferase assay (Promega) following manufacturer's protocol as described previously. The remaining half of the cell volume was kept for total RNA extraction in order to perform the RNA level quantifications by reverse transcription (as described previously), followed by the droplet digital PCR (ddPCR) quantification of the cDNA.

**ddPCR quantification**

The ddPCR quantification was performed by the RNomics Platform of the Université de Sherbrooke. Briefly, the ddPCR reactions for both Fluc and Rluc were composed of 10 µL of 2X QX200 ddPCR Supermix for probe (Bio-Rad), 10 ng (3 µL) of cDNA, a 250 nM final concentration of the probe solutions for both Fluc (FAM, from IDT) and Rluc (HEX, from IDT) and a 0.9 µM final concentration of the primer pair solutions for each target gene in a 20 µL reaction. The ddPCR fourplex reactions for the Reference genes were composed of 10 µL of 2X QX200 ddPCR Supermix for probe (Bio-Rad), 10 ng (3 µL) cDNA, a 250 nM final concentration of the probe solutions for MRPL19 (FAM) and YWHAZ (HEX), a 125 nM final concentration for the probe solutions for both PUM1 (FAM) and B2M (HEX) and a 0.9 µM final concentration of the primer pair solutions for each reference gene in a 20 µL reaction.

Each reaction mix (20 µL) was converted to droplets with the QX200 droplet generator (Bio-Rad). Droplet-partitioned samples were then transferred to a 96-well plate, sealed and cycled in a C1000 deep well Thermocycler (Bio-Rad) using the following cycling protocol: 95°C for 5 min (DNA polymerase activation); 50 cycles of 95°C for 30 sec (denaturation), 59°C for 1 min (annealing) and 72°C for 30 sec (extension); and post-cycling steps of 4°C for 5 min, 90°C for 5 min (signal stabilization) and a hold at 4°C. Reference gene reactions were cycled using the following cycling protocol: 95°C for 5 min (DNA polymerase activation); 50 cycles of 94°C for 30 sec (denaturation), 59°C for 1 min (annealing/extension); and post-cycling steps of 98°C for 10 min (enzyme deactivation) and a hold at 4°C. The cycled plate was then transferred and read using the QX200 reader (Bio-Rad) either immediately or the next day. The concentration reported is in copies/uL of the final 1x ddPCR reaction (using QuantaSoft software from Bio-Rad)(Taylor *et al.*, 2015). The Rluc and Fluc luciferases mRNA expression levels were normalized using the copies/µL ddPCR quantification, and are reported as percentages of expression relative to the WT sequence which was set to 100%.

**SHAPE and RNA secondary structure analyses**

*Transcription of RNA*

The DNA templates for the *in vitro* transcription of the RNAs to be used for SHAPE were created by the amplification of 5 ng of the psiCHECK-2 constructions with either the complete WT or the rG4mut BAG-1 5'UTR using primers that inserted the RNA polymerase T3 promoter binding site at the 5' end of the BAG-1 5'UTR and conserved the next 40 nts of the psiCHECK-2 plasmid sequence at the 3'end of the BAG-1 5'UTR. The BAG-1 5'UTR with the IRES mutated sequences were amplified from either the WT or the rG4mut pRL-BAG1-HL constructions, inserting the T3 promoter binding site upstream of the BAG-1 5'UTR and adding the same 40 nts-long 3' end flanking sequence then the constructs originating from the psiCHECK-2 plasmid. The primers used for DNA template preparation and amplification are listed in **Supplementary Table S3 in Annexe 5,** and the full RNA sequences used for SHAPE are listed in **Supplementary Table S4 in Annexe 5**. The *in vitro* transcriptions were performed following the protocol described previously (Giguère et Perreault, 2017).

*Selective 2'-hydroxyl acylation analyzed by primer extension*

The pre-folding of RNA (5 pmoles) was performed in folding buffer (20 mM Li Cacodylate pH 7.5, 100 mM KCl) in a total volume of 10 µL. The RNA was incubated for 5 min at 75°C and then slow-cooled to room temperature for 1 h. The selective 2'-hydroxyl acylation reaction was performed by adding either 1 µL of a freshly prepared 600 mM solution of the SHAPE reagent Benzoyl Cyanide (BzCN, CAS#613-90-1, Sigma-Aldrich, dissolved in DMSO) or 1 µL of DMSO (no SHAPE reagent control) and incubating for 1 min at room temperature. A volume of 90 µL of $H_2O$ was then added and the RNA was ethanol precipitated and the resulting pellet was dissolved in 10 µL of 0.5X TE buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA). Subsequently, the primer extension step was performed using the Superscript III Reverse transcriptase (Life technologies). Two primer extension reactions were performed in parallel using two different 6-FAM-labeled primers (Applied Biosystems), one for each reaction. Primer 1 bound the flanking 28 nts located at the 3' end, and primer 2 the middle of the 5'-UTR (at positions 301 to 320) in order to compensate for the reduced reverse transcription of the enzyme after ~300 nts. The RNA was unfolded by

heating at 95°C for 3 min, and was then snap-cooled on ice. Annealing of the 6-FAM labeled primers (1 pmol) was performed by heating to 65°C for 5 min; then 37°C for 5 min and finally 4°C for 1 min. The reverse transcriptase reaction was then performed for 30 min at 52°C in a buffer with final concentrations of 1X first strand buffer, 10 mM DTT, 1 mM of each dNTP and 20 % DMSO.

In order to obtain the DNA sequencing reactions necessary for the subsequent quantitative SHAPE analysis, primer extensions reactions were performed on untreated RNA sequences. The primer extension reactions to obtain the sequencing reactions were performed under the same reverse transcriptase conditions using 5 pmoles of RNA without pre-folding and in presence of an additional 1 mM of either ddCTP or ddGTP and using the corresponding NED-labeled primer 1 or 2 (Applied Biosystems). The fluorescent primers used are listed in **Supplementary Table S3 in Annexe 5**. Following the primer extension reactions for both the SHAPE reactions and the sequencing reactions, 2 µL of 2 N NaOH was added to each and the samples heated at 95°C for 5 min to degrade the RNA. The cDNA samples were ethanol precipitated and the resulting pellet air-dried. Capillary electrophoresis of the cDNA was performed at a sequencing and genotyping facility: Plateforme de séquençage et de génotypage (CHUL, Québec, Canada). There, the DNA pellets were dissolved in a mixture of 10 µL each of $H_2O$ and formamide with the addition of a Lyz labelled control DNA ladder (Life Technologies). Each SHAPE reaction and no SHAPE reagent control reaction was electrophoresed in the presence of the ddCTP sequencing reaction on an ABI 3100 Genetic Analyzer (Life Technologies). The electrophoresis was then repeated with both the SHAPE and the no SHAPE control reactions in the presence of the ddGTP sequencing reactions.

*Quantitative SHAPE analysis and secondary structure prediction*

Quantitative SHAPE reactivity for each nucleotide was determined from the electropherograms using the QuSHAPE software version 1.0 (Karabiber *et al.*, 2013). The normalized reactivity for each nucleotide was then averaged from the four SHAPE experiments (two replicates with primer 1 and two replicates with primer 2) and used as pseudo-energy constraints for RNA secondary structure prediction using the default slope (1.8 kcal/mol) and intercept (-0.6 kcal/mol) values in the Fold tool of the RNAstructure software version 5.7 (Reuter et Mathews, 2010). Comparison and clustering of the ensemble of possible secondary structures respecting the SHAPE constraints for the different RNA

sequences was performed using the StructureXplore software(Glouzon *et al.*, 2017a). The predicted minimum free energies (MFE), in kcal/mol, of the different regions of the secondary structures were evaluated using the *RNAeval* function of the ViennaRNA package (Lorenz *et al.*, 2011). The secondary structure representations were made with VARNA (Darty *et al.*, 2009) and the Arc-plots were made with R-CHIE (Lai *et al.*, 2012).

**Statistical analysis**

All statistical analysis and tests were performed using GraphPad Prism version 7.03 for Windows (GraphPad Software, La Jolla, CA, USA, www.graphpad.com). Each statistical tests performed, including the number of replicates and the number of independent experiments, are indicated in the figure legends. *P*-values < 0.05 were considered significant.

## RESULTS AND DISCUSSION

### BAG-1 expression is post-transcriptionally regulated in CRC cell lines and tumors

The rG4 located in the 5'UTR of the BAG-1 mRNA was identified during our analysis of the rG4s associated with the CRC pathway that could affect mRNA translation (Jodoin et Perreault, 2018).The initial step was thus to confirm the post-transcriptional regulation of the BAG-1 mRNA in human CRC cells. In order to do so, the expression levels of the BAG-1 mRNA and of the protein isoforms were measured in paired tissue samples extracted from colorectal tumors at different stages and from their surrounding healthy margins. These analyses were also performed in both normal and cancerous colorectal cells in culture. We initially speculated that if BAG-1 expression is indeed post-transcriptionally regulated in colorectal tumor settings, the RNA levels should not correlate with the protein isoform levels.

The RNA levels were compared in eight tissue pairs for each of the adenomas and the stage 1 tumors, and in ten tissue pairs for each tumor of stages 2, 3 and 4. With the exception of the stage 1 tumors, all of the tumor stages demonstrated statistically significant differences in the levels of the BAG-1 mRNA expressed in the tumor tissues versus the margins, with the majority of the pairs showing decreased levels of the BAG-1 RNA in the tumor (**Figure 36A**).

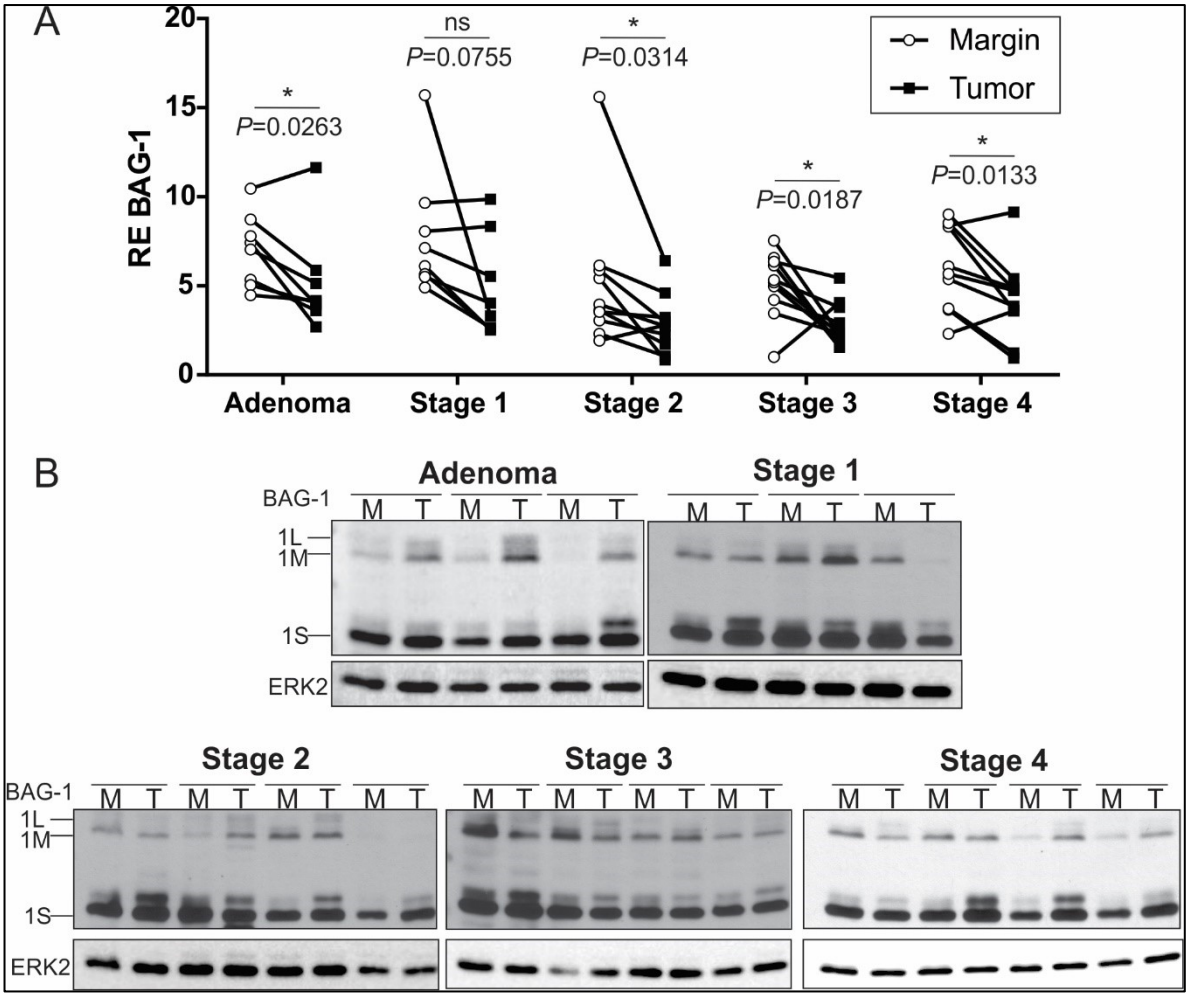



**Figure 36** – RNA and protein expression levels of BAG 1 in the paired tissues of colorectal tumors at different stages and their adjacent healthy tissue (margin).

(A) Relative expression levels of the BAG-1 mRNA from the paired adjacent healthy tissues (Margin, white circle) and from the tumors (black square) at different stages measured by RT-qPCR. Each tissue pair is connected by a line ($n$= 8 for adenoma and stage 1, $n$=10 for stages 2, 3 and 4). The statistical analysis performed is a paired t-test, tumor compared to margin, ns = no significant difference, * $P \leq 0.05$. (B) Protein expression levels, as measured by Western blot, of the three BAG-1 isoforms in the same pairs of margin-tumor tissues as in (A). ERK2 is used as the loading control ($n$=3 for adenoma and stage 1, $n$=4 for stages 2, 3 and 4). The Western blot of the remaining tissue pairs are available in the supplementary material.

Next, to measure the protein expression levels of the three BAG-1 protein isoforms, Western blots were performed using an antibody that recognizes their common C-terminal region. The protein extracts were derived from the same paired tissue samples as in the **Figure 36A**. For the three adenoma tissue pairs, an increase in all isoforms was observed in the tumor tissues (**Figure 36B**). This is also observed in the four supplementary adenoma

tissue pairs analyzed (**Supplementary Figure S1 in Annexe 5**). An increase in protein isoform expression levels, or changes in the molecular weights of the isoforms, was also observed in some of the tumor tissue samples from stages 1 to 4 (**Figure 36B** and **Supplementary Figure S1 in Annexe 5)**. The loss of expression of the longer isoform BAG-1L was also observed in one tissue-pair at stage 2. In contrast to the mRNA levels that decreased in the tumors of different stages, the protein levels either stayed constant or increased in the tumors. This absence of correlation between the mRNA and protein expression levels suggest a post-transcriptional regulation of BAG-1 expression.

This observation was also supported by the RNA and protein levels that were measured in nine cancerous colorectal cell lines and that were compared to two normal intestinal epithelial cell lines (HIEC and CRL-1831). The BAG-1 mRNA level was decreased by at least 2-fold in cancer cell lines compared to normal cells (**Supplementary Figure S2A in Annexe 5**). The protein levels of the three BAG-1 isoforms in pooled protein lysates from seven of the colorectal cancer cell lines were measured by Western blot and compared to the BAG-1 isoform protein levels in the normal HIEC cell line. Different expression levels of the BAG-1 protein isoforms were observed depending on the cell line used, but all of them presented an increase in the expression of the BAG-1 isoforms as compared to HIEC cells (**Supplementary Figure S2B in Annexe 5**). This observation is in agreement with previous work (Southern *et al.*, 2012) which also observed an increased BAG-1 protein expression levels in an array of CRC cells lines. This absence of correlation is not surprising, as proteogenomic analyses previously demonstrated that the mRNA abundance is not a good predictor of the protein abundance level in both colon and rectal tumors (Zhang *et al.*, 2014). From our observations in both colorectal cell lines and in colorectal tumors tissue pairs, the BAG-1 expression seems to be post-transcriptionally regulated as the RNA levels did not correlate with the protein levels. The presence of an rG4 in the 5'UTR of the BAG-1 mRNA transcripts might be involved in this lack of correlation.

**Stabilization of the rG4 using chemical ligands decreases BAG-1 protein isoforms expression**

BAG-1 expression is post-transcriptionally regulated in CRC cell lines and tumor samples. As rG4 are generally described as being translational repressors, the folding of the rG4 structure in the BAG-1 5'UTR could affect the expression of the BAG-1 protein isoforms.

To verify this assumption, CRC cells (HCT116) were treated with the RNA quadruplex specific ligand cPDS (Di Antonio *et al.*, 2012). The binding of the ligand should stabilize the rG4 structure. As most rG4s located in the 5'UTR impede translation due to steric hindrance caused by their high stability, further stabilization of the structure upon ligand binding should enhance this repressive effect (Bugaut *et al.*, 2010; Gomez *et al.*, 2010). Upon treatment with 2 μM cPDS, the protein levels of the three isoforms decreased by almost 2-fold (**Figure 37A and B, left panel**). Nevertheless, the BAG-1 RNA levels were unchanged after treatment (**Figure 37C**), demonstrating that the repression happened at the post-transcriptional level. To eliminate the possibility that this result was due to a cPDS treatment artefact, two other ligands specific for G4 were also used: Phen-DC3 and TMPyP4. Both these ligands are known to specifically bind and stabilize DNA G4 (Cian *et al.*, 2007; Izbicka *et al.*, 1999). However, the TMPyP4 ligand has an opposite effect on RNA G4 structures as it has been described to destabilize and unfold RNA quadruplexes (Morris *et al.*, 2012). In accordance with the opposite stabilization effects of these two ligands, treatment of the cells with 10 μM of Phen-DC3 resulted in a 2-fold decrease in the protein isoform levels, and treatment with 2 μM of TMPyP4 had no effect the protein expression levels (**Figure 37B, middle and right panel**). Identical to the cPDS treatment, the RNA levels were unchanged by treatment with the two supplementary ligands (**Figure 37C**). This assay demonstrated that the stabilization of the rG4 located in the 5' end of the BAG-1 UTR decreased the protein expression levels of the three downstream isoforms, and that this effect is post-transcriptional.

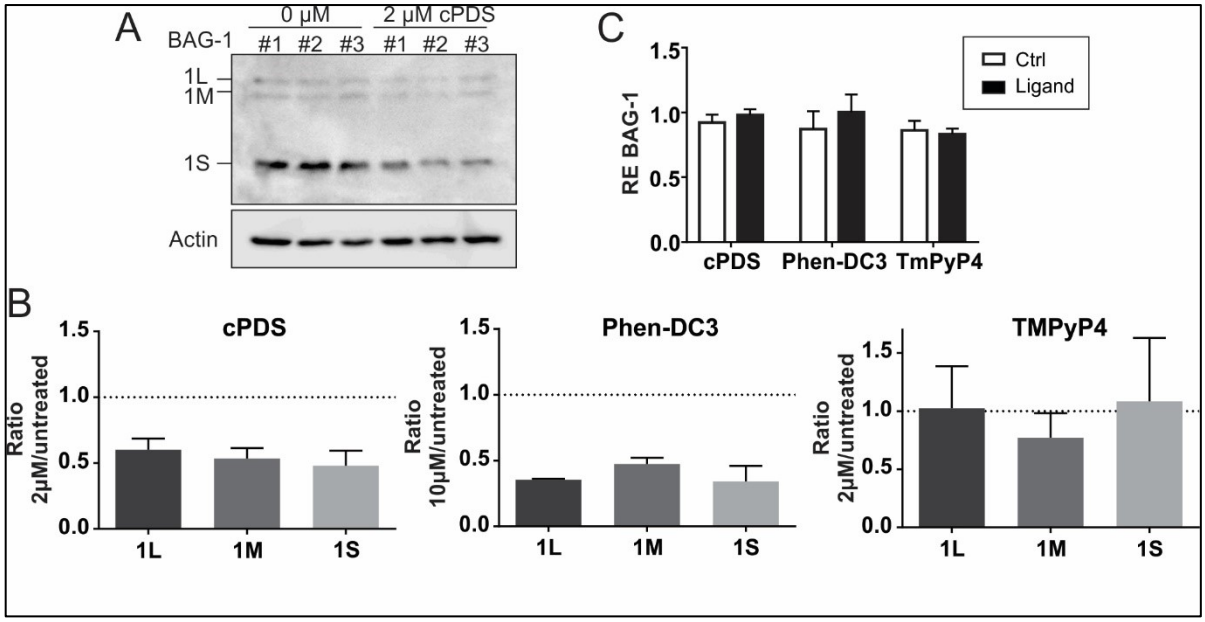



**Figure 37** – Stabilization of the rG4 with chemical ligands.

(A) A representative gel of the BAG-1 protein isoform levels after a 24 h treatment with 2 µM of the specific rG4 ligand cPDS compared to that of an untreated control. (B) Ratios of the BAG-1 isoform protein levels of the ligand treated cells over the untreated cells. (C) Relative expression of the BAG-1 mRNA in the ligand treated cells compared to the untreated cells. For both (B) and (C), the results are presented as the means and standard deviations of $n$=2.

**Disruption of rG4 formation through mutations increases reporter gene expression**

Stabilization of the rG4 by small-molecule ligands resulted in a decrease of the protein isoform's expression levels. However, it should be noted that these ligands can target all rG4s present in the cells and not the BAG-1 rG4 specifically. The next step was thus to directly mutate the rG4 using G-to-A mutations that abolish all potential for rG4 formation and then measure the effect on both the RNA and protein expression levels.

The most common method for assessing any rG4's effect on the expression level of a given mRNA *in cellulo* is to insert the complete 5'UTR upstream of a luciferase reporter gene and to compare the expression level to that of a second construction in which the rG4 is mutated. Using this technique, it was demonstrated that the abolition of the rG4 located in the BAG-1 5'UTR resulted in a 3-fold increase in luciferase expression in HCT116 cells as well as in two other CRC cell lines (HT-29 and DLD-1) (Jodoin et Perreault, 2018). However, a limitation of this previous work was that only the 5'UTR region corresponding to positions 1 to 87, upstream of the CUG start codon was used. Here, the experiment was repeated using

the complete 501 nts of the 5'UTR that includes all of the alternative start codons that are located downstream of the rG4. With this complete 5'UTR construction, mutation of the rG4 resulted in a 1.7-fold increase in the luciferase expression level even when the 1S start codon coding for the most abundant of all three isoforms was mutated from AUG to AGG (**Figure 38A**).

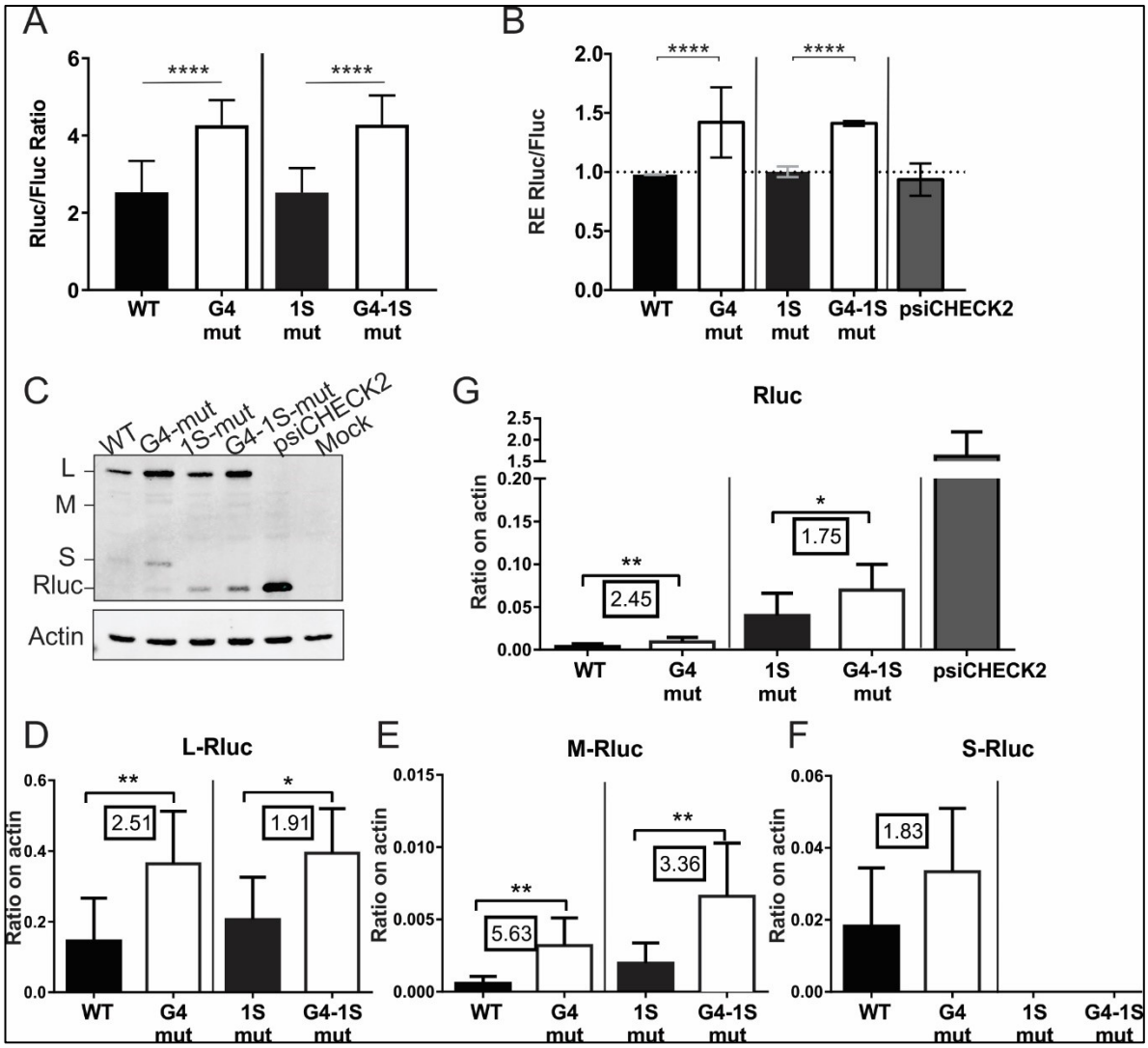



**Figure 38** – Luciferase, RNA and protein isoform expression levels from reporter assays of the complete 5'UTR of BAG 1 with both the mutated rG4 and the mutated 1S start codon.

(A) The luciferase assays' means and standard deviations of the Rluc luminescence levels, normalised over the Fluc normalisation levels, are shown for the WT (black) and rG4 mut (white) psiCHECK-2 constructions that included or not the 1S start codon mutation. The experiments were repeated three times with each of the constructions being transfected in triplicate ($n$=3). The statistical analysis performed is a two-way ANOVA with Tukey's multiple comparison, ****$P$≤0.0001. (B) Relative

expression (RE) levels of the Rluc RNA normalised over that of the Fluc RNA after the transfections of the different mutated constructions measured by RT-qPCR. The bar of the RE level of the reporter plasmid without the insertion of the BAG-1 5'UTR is labeled psiCHECK-2. The statistical analysis performed is a two-way ANOVA with Tukey's multiple comparison test ($n$=2), ****$P$≤0.0001. (**C**) Representative immunoblot of the Rluc N-extension protein isoforms' expression levels after both the rG4 and 1S start codon mutations. The psiCHECK-2 transfection lane represents the canonical Rluc without any N-terminal extension. Mock indicates the untransfected control. β-actin was used as a loading control. (D-G) Quantification of the level of each isoform, normalised over that of the β-actin loading control, (D) L-Rluc (E) M-Rluc (F) S-Rluc (G) Rluc. The boxed values are the fold-change in protein level of the rG4mut construction over that of the WT. The statistical analysis performed is a Mann-Whitney test ($n$=3). *$P$≤0.05, **$P$≤0.001, ***$P$≤0.0005.

**The rG4 affects the protein abundance of all N-terminal extension isoforms**

The presence of the three alternative start codons, all of which are in frame with the *Renilla* luciferase (Rluc) reporter gene in the complete 501 nts 5'UTR construction, could result in luciferase protein isoforms with alternative N-terminal extensions. These additions could affect the DNA reporter's transcription and both the resulting Rluc protein's folding and enzymatic activity. Consequently, in order to accurately measure the luciferase expression levels in the presence of the N-terminal extensions, immunoblots against the C-terminal region of the Rluc were performed along with RNA quantifications (**Figure 38B and C**). To check if rG4 formation affects all protein isoforms similarly, the rG4 mutation was individually combined with the mutation of each start codon. The 1L CUG start codon was mutated to CGG, and both the 1M and 1L AUG start codons were mutated to AGG. The rG4 effects were thus measured as the differences in the remaining possible isoform levels between the WT and the rG4 mutated constructions.

As anticipated, the transfection of the complete 5'UTR reporter constructions resulted in the use of the alternative start codons for translation of the Rluc reporter. Rluc isoforms with N-terminal extensions that migrated at corresponding higher molecular weights than the canonical Rluc were observed (**Figure 38C**). The canonical Rluc control (36 kDa) was seen in the psiCHECK-2 vector without insert lane (**Figure 38C**, lane 5). The dominant expression from the BAG-1 WT 5'UTR reporter construct was that of the L-RLuc isoform, along with faint M-Rluc, S-Rluc- and canonical Rluc expression levels (**Figure 38C**). The abolition of the rG4 resulted in an increased abundance of 1.8- up to 5.6-fold of all of the Rluc N-terminal extension isoforms as compared with the WT protein levels normalised to the actin loading control (**Figure 38D-G**). The 1S-mut construction, in which the start codon of the shortest isoform was abolished, resulted in the loss of protein expression of that

isoform in favor of the next start codon in the sequence, namely the canonical Rluc (**Figure 38C**, lane 3). The combination of the 1S-mut and the rG4-mut reporter construct and to the 1S-mut alone, resulted in a similar fold-increase in the protein isoform levels as was observed for the rG4-mut alone compared to the WT (**Figure 38D-G**). This demonstrated that even if the three isoforms are not equally expressed, the rG4 represses the translation of all of them to the same extent.

That said, the Rluc RNA levels were slightly different upon the transfection of the different constructions (**Figure 38B**). All constructions bearing the rG4 mutation had relative RNA expression levels 1.5-fold higher than did the constructions with the intact rG4. This increase in the RNA levels was still lower than the average 2-fold increase in the protein levels that was observed for all of the rG4-mut constructions. Thus, the increase in the protein expression levels is not directly proportional to the RNA levels. The protein levels could also be increased by a greater translation of the rG4mut constructions. Both of these effects of the rG4 on the Rluc RNA and protein isoform levels were also reciprocated using constructions in which either the 1L or the 1M start codons were mutated (**Supplementary Figure S3 in Annexe 5**). In summary, rG4 formation has a repressive effect on protein expression levels of all in-frame protein isoforms.

This result is reminiscent of both the leaky scanning and the alternative translation initiation mechanisms. At the initiation step, the 43S ribosomal complex scans the 5'UTR until it recognizes one of the in-frame start codons by complementarity, a processed that is favored by the strength of the initiation context. Because the rG4 is located upstream of all of the start codons, it impairs the scanning efficiency from the very beginning before any "encounter" with a potential start codon, thus affecting all isoform's translation. The proximity of the rG4 to the 5'-cap is also a mechanism that could explain its repressive effect on translation. rG4 that are located close to the cap are more detrimental than ones located further downstream (Kumari *et al.*, 2008) as they can impede either the co-transcriptional 5'cap synthesis or its recognition by the translation initiation factors. The BAG-1 rG4 is located very close to the 5' end, specifically at position 6. However, these hypotheses were refuted because no difference was observed between the WT and the rG4 mut BAG-1 sequences during both *in vitro* cap-synthesis assays and affinity binding assays with the cap-binding protein eIF4E (data not shown).

**A repressive uORF is present in the BAG-1 5'UTR**

In addition to the presence of multiple in-frame start codons located downstream of the rG4, the BAG-1 5'UTR also possesses start codons in the other frames. One of them, the out-of-frame AUG located at position 254, stands out as it presents a more favorable context for translational initiation than all of the in-frame start codons (**Supplementary Figure S4A and Supplementary Table S2 in Annexe 5**). The analysis of publicly available ribosome profiling data (Crappé *et al.*, 2015; Michel *et al.*, 2014; Zhang *et al.*, 2017) demonstrated the presence of ribosome-protected fragments (RPF) corresponding to initiating ribosomes at this position in different cell lines (**Supplementary Figure S4B in Annexe 5**). A UGA stop codon is located downstream at position 302 of the 5'UTR, and it could result in a short open-reading frame (ORF) of 16 amino acids. The presence of a short out-of frame ORF located upstream of the main protein coding sequence corresponds to the definition of an upstream ORF (uORF). uORF are *cis* regulatory elements that repress translation. They act as decoys for the ribosomes in order to initiate translation early, before the main ORF. Their presence creates new requirements of translational re-initiation in order to translate the main ORF (Calvo *et al.*, 2009). Well-known examples of repressive translation regulation by uORFs in 5'UTRs are the ATF4 and C/EBPα-β transcripts (Calkhoven *et al.*, 2000; Vattem et Wek, 2004). The impact of this previously uncharacterized possible uORF on the BAG-1 regulation of translation was investigated both in the presence and the absence of the rG4 structure.

To first confirm whether or not the possible uORF affects the protein expression levels, the AUG located at position 254 was mutated to ACG in the reporter gene with the full-length BAG-1 5'UTR sequence in-frame with the Rluc coding sequence. This silent mutation was chosen in order to conserve the same coding histidine in the main frame of the Rluc N-terminal extension protein isoform while completely disrupting both the start codon and the translation initiation context sequence of the possible uORF (**Supplementary Table S2 in Annexe 5**). The luciferase expression level of the mutated uORF construct was 2-fold higher than that of the WT 5'UTR construct (**Figure 39A**), indicating that this AUG-254 acts as a repressor element. At the protein level, the mutation of the AUG-254 resulted in an increase of the abundance of the M-Rluc isoform (**Figure 39B, C**). This was expected as the AUG start codon of the M-isoform located at position 301 is the next one in line after the AUG-

254 in the scanning of the 5'UTR. A slight decrease in the 1L isoform level is observed when the AUG-254 is mutated, but this difference is not statistically significant. The 1M- isoform possesses a stronger translation initiation context than the 1L-isoform. Without the repressive AUG-254 start codon, the downstream 1M isoform might be favored over the 1L. The level of the downstream 1S-isoform is also slightly increased upon the AUG-254 mutation, but the difference is statistically significant only between the rG4mut construct compared to the combined rG4mut-AUG-254mut construct.

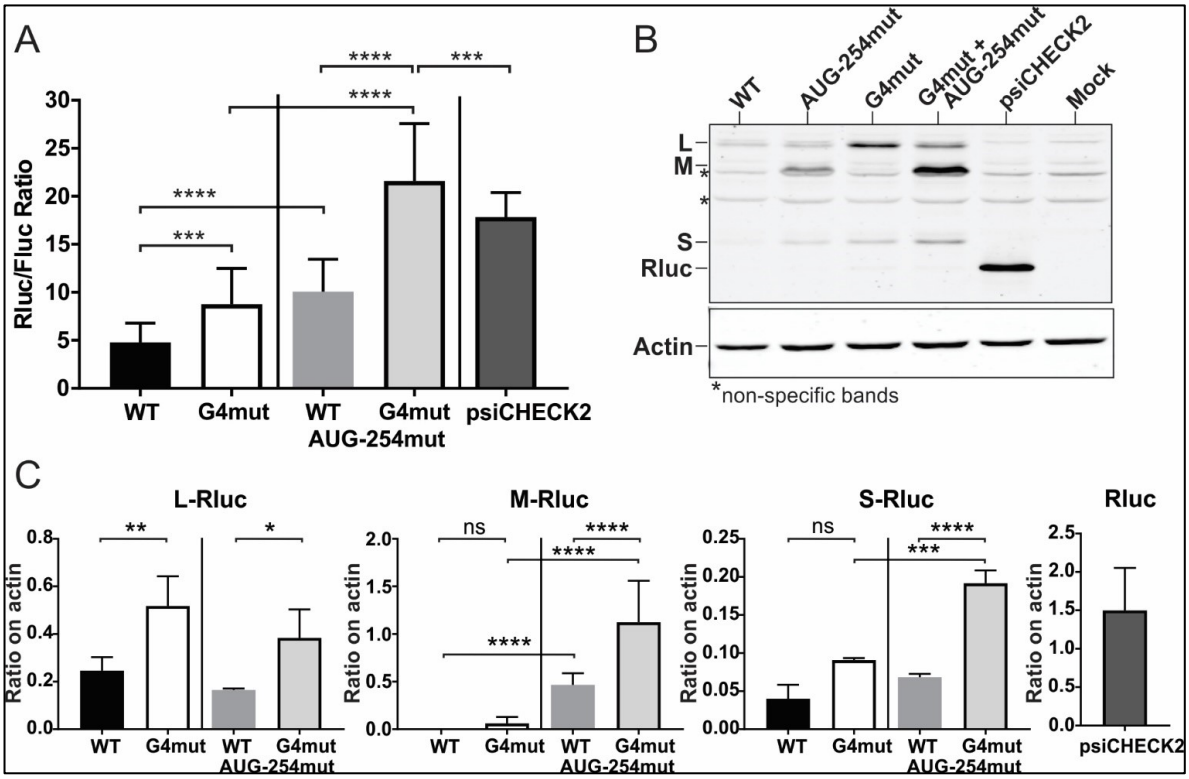



**Figure 39** – Luciferase and protein expression levels from reporter assays of the 5'UTR of BAG-1 possessing the mutated AUG-254.

(A) The luciferase assays' means and standard deviations of the Rluc luminescence levels, normalised over the Fluc luminescence levels, are shown for all constructions). The statistical analysis performed is a two-way ANOVA with Tukey's multiple comparison, ($n$=2, each construction transfected in triplicate). ***$P$≤0.001, ****$P$≤0.0001. (B) Representative immunoblot of the Rluc N-extension protein isoforms' expression levels after both the rG4 and the AUG-254 start codon mutations. The psiCHECK-2 transfection lane represents the canonical Rluc without any N-terminal extension. Mock indicates the untransfected control. β-actin was used as a loading control. (C) Quantification of the level of each isoform, normalised over that of the β-actin loading control. The statistical analysis performed is a two-way ANOVA with Tukey's multiple comparison ($n$=2, each construction transfected in triplicate), ns=not statistically significant, *$P$≤0.05, **$P$≤0.01, ***$P$≤0.001, ****$P$≤0.0001.

Upon rG4 abolition, the mutation of the AUG-254 resulted in a doubled increase of the luciferase expression level as compared to that of the AUG-254 mutation alone (**Figure 39A**). The luciferase expression is even higher than that of the psiCHECK-2 reporter control without the inserted 5'UTR. This effect is also seen at the protein level. As shown previously, the rG4 mutation resulted in an increased abundance of all of the isoforms. The combination of both rG4 and AUG-254 mutations also resulted in a doubled protein level as compared to that of the AUG-254 mutation alone (**Figure 39B, C**). As it is the case for the in-frame start codons, the rG4 also seems to repress the scanning during the very first steps of translational initiation, before the encounter with the repressive uORF, and also affects the initiation at this out-of-frame AUG in a manner similar to that seen at the in-frame start codons.

This newly characterized uORF might regulate the BAG-1M isoform, and this might explain why this isoform is expressed less than the 1L isoform despite having a canonical start codon. Because of its more favorable Kozak context, initiation is favoured at the AUG-254, rather than at the BAG-1M start codon. Furthermore, the 1M start codon located at position 301 is hidden inside the uORF sequence, which limits the chances of re-initiation and thus reduces its expression. Under stress conditions, where re-initiation is slowed down due to the reduced availability of both the ternary complex and the initiation factors, both the BAG-1M translation, as well as the BAG-1S isoform translation with the start codon situated further downstream at position 501 might be favored. This is a mechanism that is common to other transcripts that possess alternative in-frame start codons along with uORFs (Hinnebusch *et al.*, 2016). However, this remains to be validated experimentally for BAG-1. It is unknown if the uORF located at position 254 is readily translated into a short peptide, or if its only function is to divert the pre-initiating ribosome complex from the main reading frame of the BAG-1 isoforms.

The BAG-1 5'UTR possess two *cis*-elements that affect the cap-dependent translation: an rG4 and an uORF which both act additively to repress translation of the N-terminal extensions' protein isoforms. A recent study by the Balasubramanian group demonstrated that, at the genome level, rG4s are enriched in 5'UTRs with possible repressive uORFs and could stimulate translational initiation at these uORFs (Murat *et al.*, 2018). The BAG-1

5’UTR situation is very similar to their presented model, and our results fit with their hypothesis.

**Disruption of the rG4 formation is detrimental to the IRES-dependent translation**

Disruption of the rG4 formation in the BAG-1 5’UTR probably facilitated the scanning of the 5’UTR, and thus it affected the alternative translational initiation of the three principal protein isoforms. However, leaky scanning and alternative translational initiation are not the only mechanisms regulating the translation of the BAG-1 mRNA. The BAG-1 5’UTR also possess an IRES secondary structure (Coldwell *et al.*, 2001). With the collaboration of ITAFs, the secondary structure allows for both the direct recruitment of the 40S ribosomal subunit at the RBS and the initiation of translation in a cap-independent manner (Pickering *et al.*, 2004). Some rG4 prone sequences were previously found to affect both the secondary structure’s folding and the cap-independent translation of IRES. Therefore, the possible impact of the BAG-1 rG4 on the IRES-driven translation was investigated.

To detect IRES-dependent translation, bicistronic luciferase reporter DNA backbone vectors were used. In these constructions, the complete 501 nts 5’UTR of BAG-1 was inserted between the Rluc and the Fluc reporter genes. The Rluc is expressed following cap-dependent translation, while the Fluc is translated following cap-independent internal initiation of translation. The Fluc/Rluc ratio thus represents the IRES activity. Again, the WT BAG-1 5’UTR was compared to the rG4 mutant in order to observe the difference in the Fluc normalised expression levels. The well-characterized IRES from Hepatitis C virus (HCV), specifically the initial pRL-HL construction, was used as the positive control for the IRES-dependent translation (**Figure 40A**). In opposition to the monocistronic luciferase construction where the rG4 mutation triggered a higher luciferase expression, the transfection of the rG4 mutated bicistronic construction produced a small but consistent decrease of 20% in the Fluc/Rluc ratio (**Figure 40A**). The effect is translational as no difference is observed in the RNA expression levels of the constructions (**Figure 40B**). The absence of monocistronic Fluc products resulting either from cryptic promoter usage or unexpected splicing was confirmed by Northern blot analyses using both Rluc and Fluc specific probes (**Supplementary Figure S5 in Annexe 5**).

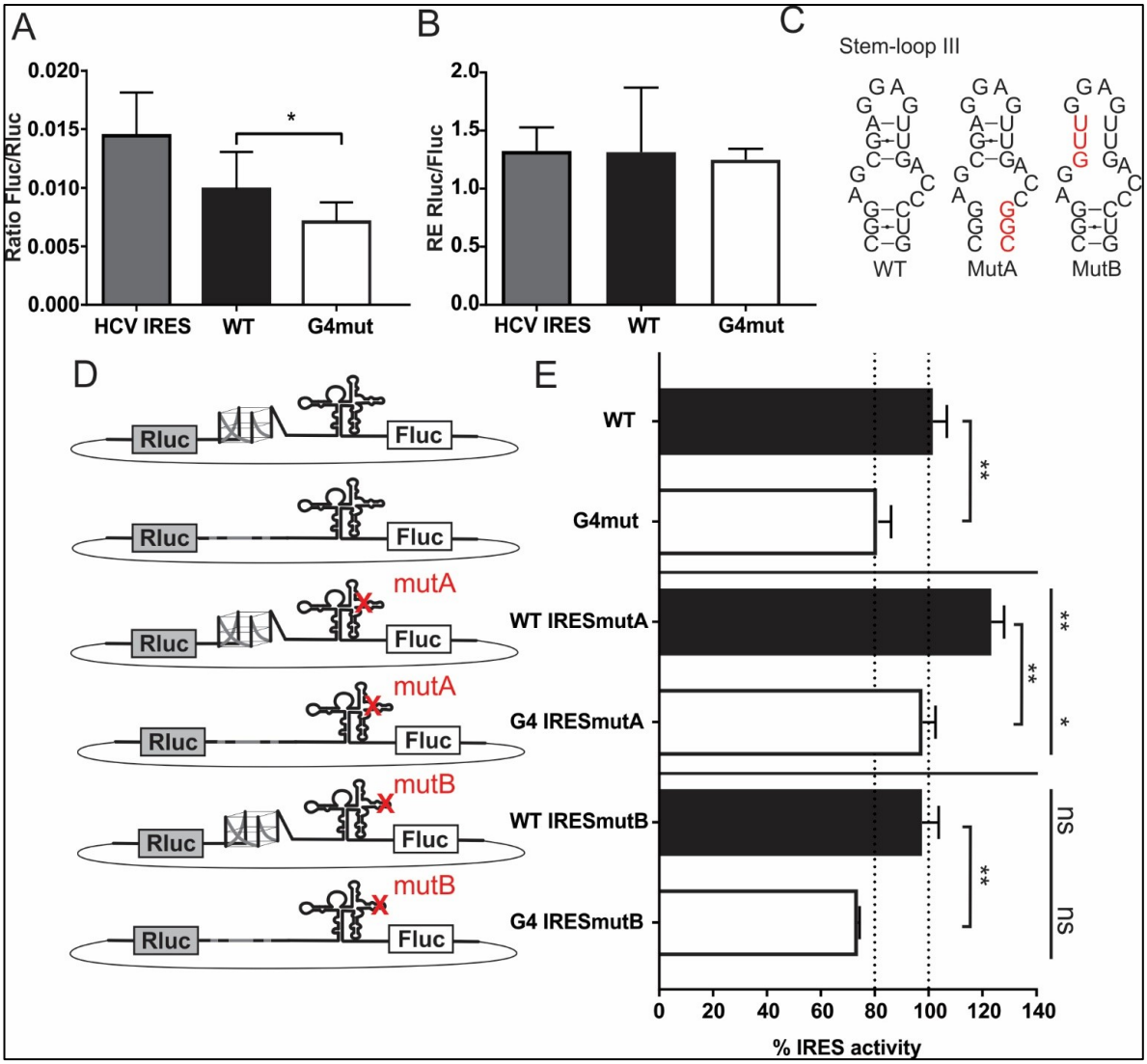



**Figure 40** – rG4 mutation impairs the cap-independent translation of the BAG 1 IRES.

(A) The ratios of Fluc/Rluc luciferase levels following transfection of the bicistronic plasmid construct are shown. The HCV IRES (gray) bicistronic construct was used as a positive control, with the well characterized IRES placed upstream of the Fluc. The WT (black) represent the bicistronic construct with the full length 5'UTR of BAG-1 located upstream of the Fluc. The G4mut (white) represents the bicistronic construct with the full-length 5'UTR of BAG-1 that included G-to-A mutations abolishing the folding of the rG4. For each experiment, all constructions were transfected in triplicate. The results are the means and standard deviations of $n$=5 independent experiments. The statistical analysis performed is a paired t-test, * $P \leq 0.05$. (B) The ratios of the relative RNA expression levels of Rluc and Fluc following transfection. The ratios are close to 1 and are similar between the three constructs, demonstrating the integrity of the bicistronic construct. The bars indicate the means and standard deviations of $n$=3. (C) Representation of the Stem-loop III secondary structure as defined by (Pickering *et al.*, 2004) with the IRESmutA, containing the GUC to GCC mutation at positions 367 to 369, and the IRESmutB, containing the CGA to GUU mutation at positions 354 to 356. (D) Schematic representations of the bicistronic plasmid constructions with the various rG4 and IRES structure mutations used in the assays. (E) Percentage of IRES activity for each

construct. The 100% activity level was defined as the Fluc/Rluc ratio of the WT construct. WT constructions in which the rG4 is intact are in black; the rG4mut constructions, in which the rG4 is abolished by G/A-mutations, are in white. The bars represent the means of two assays ($n$=2), each sample was transfected in triplicate, and the error bars represent the standard deviations. The top horizontal bar represents the statistical significance as compared to the IRESwt constructions (WT or G4mut, respectively). The statistical analysis performed is a one-way ANOVA with Tukey's multiple comparisons * $P \leq 0.05$, ** $P \leq 0.01$.

A decrease in the IRES-activity of only 20% was considered as low, so comparisons of this reduction with those of the other mutations known to affect the BAG-1 IRES activity were performed. Pickering *et al.* (Pickering *et al.*, 2004) deciphered the secondary structure of the minimal IRES region of BAG-1 (corresponding to positions 247 to 432 of the 5'UTR) and identified the stem-loop III as being essential for both the recruitment of ITAFs and the IRES-dependent translation (**Figure 40C**). In their work, the mutation of either the bottom part of the stem-loop III (MutA), or the upper part (MutB), significantly reduced the IRES activity in an *in vitro* translation assay using rabbit reticulocyte lysate. Those mutations were thus added to the bicistronic constructions and compared to both the BAG-1 WT and the rG4 mutant (**Figure 40D**). Surprisingly, the introduction of these IRES mutations in either the WT or the rG4mut bicistronic constructions did not reduce the IRES activity. The IRESmutB presented an IRES activity identical to that of the WT construction, while the IRESmutA resulted in a 20% increase in the IRES activity (**Figure 40E**). Independently of the presence of either IRESmutA or IRESmutB, the rG4 mutation still resulted in a 20% decrease in the IRES activity as compared to that of the corresponding intact rG4 construction.

The discrepancies in the IRES activity levels following the Stem-loop III mutations observed in both the initial work of Pickering *et al*. and this work could be explained by the different translational systems used, specifically rabbit reticulocyte lysates initially and the transfection in HCT116 cells here. Furthermore, the initial sequence for the IRES secondary structure determination and translation assays did not include the nucleotides of the rG4 region (positions 6 to 35). Hence, it is possible that the rG4 secondary structure folding impacts the global secondary structure folding of the 5'UTR, influences secondary structure long-range interactions or allows for the folding of an alternative secondary structure that could affect the IRES efficiency and mitigates the IRES mutations A and B. Although this assay did not provide a complete negative control of IRES activity for comparison, it did demonstrate that the 20% decrease in the expression of the DNA bicistronic luciferase

transfection assay was reproducible, and therefore that the abolition of the rG4 affected the IRES-dependent translation in a lesser, but opposite, way as compared to that of the cap-dependent translation.

**Cap-dependent translation is the main translational mechanism of the BAG-1 5'UTR under normal growth conditions**

The DNA transfection of the monocistronic luciferase construct containing the complete BAG-1 5'UTR demonstrated that the rG4 repressed expression because its abolition increased the level of luciferase protein (**Figure 38C**). In this assay, the mRNA levels of the rG4mut constructs were also slightly increased as compared to the constructions with the intact rG4 region (**Figure 38B**). In the DNA transfection of the bicistronic constructs, where the complete 5'UTR was located between the two luciferases, the rG4 had the opposite effect: its abolition resulted in a reproducible 20% decrease in the IRES activity (**Figure 40E**). Multiple controls were performed to eliminate the possibility of artifacts resulting from *in cellulo* modifications of the bicistronic DNA construct after transfection. Nevertheless, to limit the differences observed in the RNA levels between the WT and rG4mut monocistronic constructions in the initial transfections, and to directly account for differences at the translational level, the luciferase reporter assays were repeated using the direct transfection of exact amounts of monocistronic and bicistronic capped and poly-adenylated luciferase reporter mRNAs. Furthermore, the assays were normalized on the RNA levels post-transfection using the reverse transcription of the total RNA extracts and the ddPCR quantification of the resulting cDNA for both the monocistronic and bicistronic mRNA constructs.

All of the mRNA construct templates were created using the BAG-1 5'UTR bicistronic DNA vector and different set of primers (see **Methods**). The resulting templates were then *in vitro* transcribed, capped with either the canonical $m^7$G-cap or the A-cap analog, and then polyadenylated. The monocistronic mRNA constructs bearing the BAG-1 5'UTR upstream of the Fluc reporter coding sequence were co-transfected with the Rluc monocistronic control (**Figure 41A**). The Fluc expression level was normalised over the Rluc expression level (Fluc/Rluc ratio) for each construction, either mono- or bicistronic and was corrected by the RNA levels. They were further plotted as the relative expression level as compared to the WT monocistronic construct which was set to 100% (**Figure 41B**).

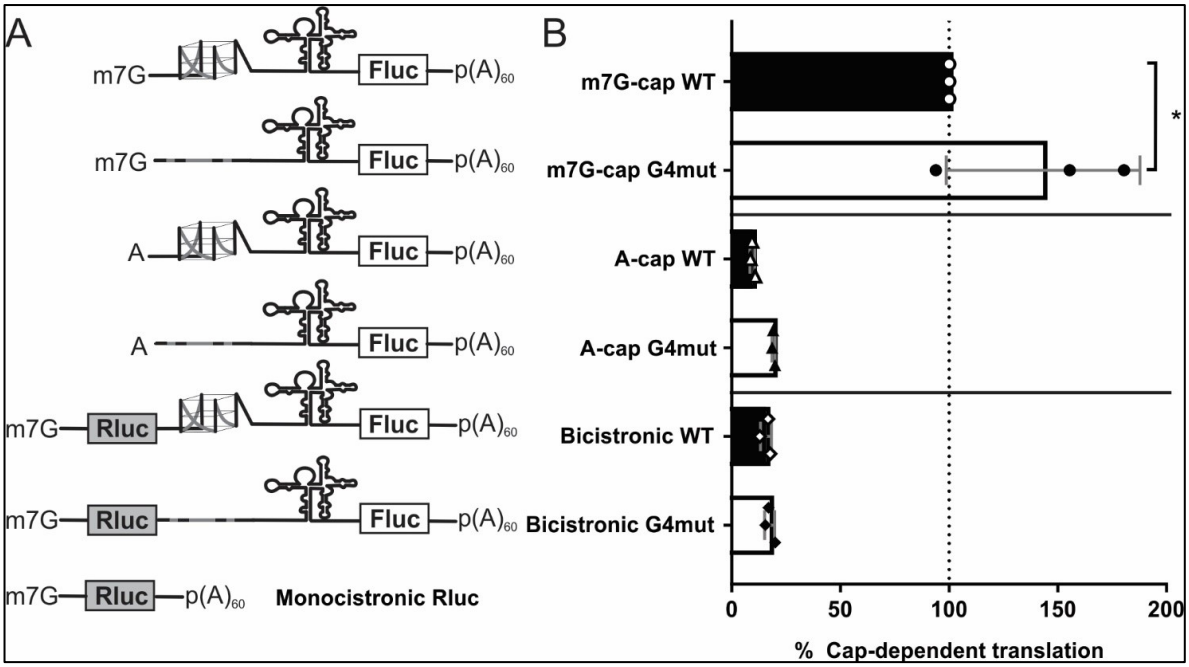



**Figure 41** – Effects of the rG4 on both the cap-dependent and the cap-independent translation of the transfected mRNA reporter constructions.

(A) Schematic representation of the mRNA constructions used in the assay. They differ first by being either monocistronic (Fluc only) or bicistronic (Rluc cap-dependent and Fluc cap-independent) and second, by the presence of either the canonical $m^7G$-cap or the analog A-cap. The monocistronic Rluc mRNA serves as the co-transfection control for the Fluc/Rluc normalisation of the monocistronic constructions. (B) The percentage of cap-dependent translation for each construct. The 100% level was set as the Fluc/Rluc ratio of the luciferase expression levels of the the WT monocistronic construction corrected by the RNA expression level as measured by RT-ddPCR. The assay was repeated three times with each construction transfected in triplicate. Each data point is the mean of the triplicate luciferase expression levels normalised over RNA expression levels. ($n$=3). The statistical analysis performed is a one-way ANOVA with Sidak's multiple comparisons test, *$P \leq 0.05$.

The monocistronic mRNA transfection repeated the effect of the rG4 observed in the first DNA transfection luciferase assay: the mutation of the rG4 resulted in an increase in translation (**Figure 41B**). In the presence of the A-cap analog, which controlled for the 5'end-dependent but $m^7G$-independent translation, the translation level of the WT 5'UTR was significantly lower at 9.5% of the $m^7G$-dependent translation. The mutation of the rG4 seemed to increase translation up to 19%, however, the difference was not statistically significant due to the low translation level. This indicates that the rG4 could repress both the $m^7G$- and 5'end-dependent translational mechanisms. For the bicistronic mRNAs, no difference in the translational levels was observed between the WT and the rG4mut, with translation levels corresponding to 15.8% and 17.5% of those of the cap-dependent

translation, respectively. The decrease in the IRES activity upon rG4 mutation was not observed in this case. In this assay, the bicistronic translation levels were so low, as compared to that observed with the $m^7$G-cap monocistronic mRNAs that a 20% reduction might be impossible to detect. By comparing the Rluc/Fluc expression levels of the different monocistronic and bicistronic mRNA constructions, it is clear that the dominant translation mechanism of the BAG-1 isoforms is cap-dependent in the normal HCT116 growth conditions used here. This is consistent with previous studies that indicated that the IRES-dependent translation occurs under stress conditions (Coldwell *et al.*, 2001; Dobbyn *et al.*, 2008). However, the initially observed 20% repression of the IRES-dependent translation that occurs when the rG4 is mutated could be explained by the impact of the rG4 on the global 5'UTR folding affecting the stability of key subdomains of the IRES secondary structure.

**Formation of the rG4 affects the global 5'UTR's secondary structure**

Stable secondary structures located near the $m^7$G-cap are known to impede both the scanning of the ribosome and the initiation of translation (Babendure *et al.*, 2006). Furthermore, structural accessibility of the regions surrounding the start codons also affects the translational efficiency, and can influence the leaky scanning mechanism (Corley *et al.*, 2017). The cap-independent translation mechanism is also very dependent on the accurate secondary structure folding of the IRES. The secondary structure folding of the 5'UTR is thus important for both types of translation initiation, and the impact of the rG4 on the global 5'UTR secondary structure might explain its apparent opposite effects on the cap-dependent and -independent translation mechanisms.

To investigate whether or not the rG4 abolition could affect the global 5'UTR folding, and more specifically the secondary structure surrounding each of the start codons and the IRES secondary structure, selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) was performed on the complete WT, the rG4mut and the IRESmutA BAG-1 5'UTR in the presence of 100 mM KCl. In each construction, a 40-nts extension was added to the 3'end of the 501 nts *in vitro* transcribed RNA 5'UTR in order to allow for the primer binding required for reverse transcription. A second primer, binding in the middle part of the UTR (positions 301 to 320) was also used in the primer extension step to optimize the reverse-transcriptase coverage of the 501 nts. The cDNAs were then analyzed using capillary electrophoresis. Flexible nucleotides from the secondary structure are more prone to react

with the acylating SHAPE reagent, creating more stops at those positions during primer extension. The averaged reactivity of each nucleotide, from two independent SHAPE experiments from each primer, was used as pseudo-energy constraints in order to predict the secondary structure using the RNAstructure algorithm (Reuter et Mathews, 2010). This software cannot predict rG4 secondary structures. Hence, to avoid base-pair predictions for the guanines of the G-tracts elsewhere in the UTR, predictions were also performed with the constraint that the nucleotides located at positions 1 to 35 remain single-stranded (G4ss). Up to 18 possible secondary structures respecting both the SHAPE pseudo-energy and the G4ss constraints were obtained for each 5'UTR WT and for the mutated sequences (**Table 8**).

**Table 8** Number of secondary structure predictions generated by RNAstructure for each of the mutated sequences using the SHAPE pseudo energy constraints

| | **Number of secondary structure predictions** | | |
|---|---|---|---|
| **Sequences** | **Pseudo-energy constraints only** | **Pseudo-energy constraints + G4ss[1]** | **Total** |
| **WT** | 8 | 17 | 25 |
| **G4mut** | 12 | 14 | 26 |
| **WT IRESmutA** | 16 | 13 | 29 |
| **G4mut IRESmutA** | 16 | 18 | 34 |
| **Total** | 52 | 62 | 114 |

1. G4ss represent secondary structure predictions in which the G4 region was constrained to stay single stranded

The StructureXplor software was then used to compare and cluster similar secondary structures (Glouzon *et al.*, 2017a). This software uses the combinations of short secondary structure motifs (Super-n-motif), instead of sequence alignment, to assess secondary structure similarities. Thus, it can compare secondary structures obtained from different mutated sequences (Glouzon *et al.*, 2017b). The ensemble of the possible predicted structures from the WT, the rG4mut and the IRESmutA 5'UTR sequences could be separated into three distinct secondary structure clusters of different sizes with the cluster 3 containing less structures than clusters 1 and 2 (**Figure 42A**). The quality of the clustering was evaluated using the computed silhouette coefficient (possible values from -1 to 1, 1 being the highest clustering quality) giving values of 0.616; 0.772; 0.711 for clusters 1 to 3, respectively. This signifies that the secondary structures are similar within each cluster, and that they are well-

differentiated between the different clusters. Of note, the secondary structures of the WT and the rG4mut sequences were not uniformly distributed in the 3 clusters. The predicted structures of the WT sequences are mostly regrouped in cluster 1, while the predicted structures of the rG4mut sequences are in cluster 2 (**Figure 42B,** top). This demonstrated that, globally, the predicted secondary structures ensemble obtained is different between the two. The abolition of the rG4 folding results in the alteration of the global secondary folding of the 5'UTR. However, the IRES mutation A does not affect the global folding, as sequences bearing that mutation are clustered in the same secondary structure ensemble proportions as are the WT or the rG4mutation alone (**Figure 41B**, bottom).

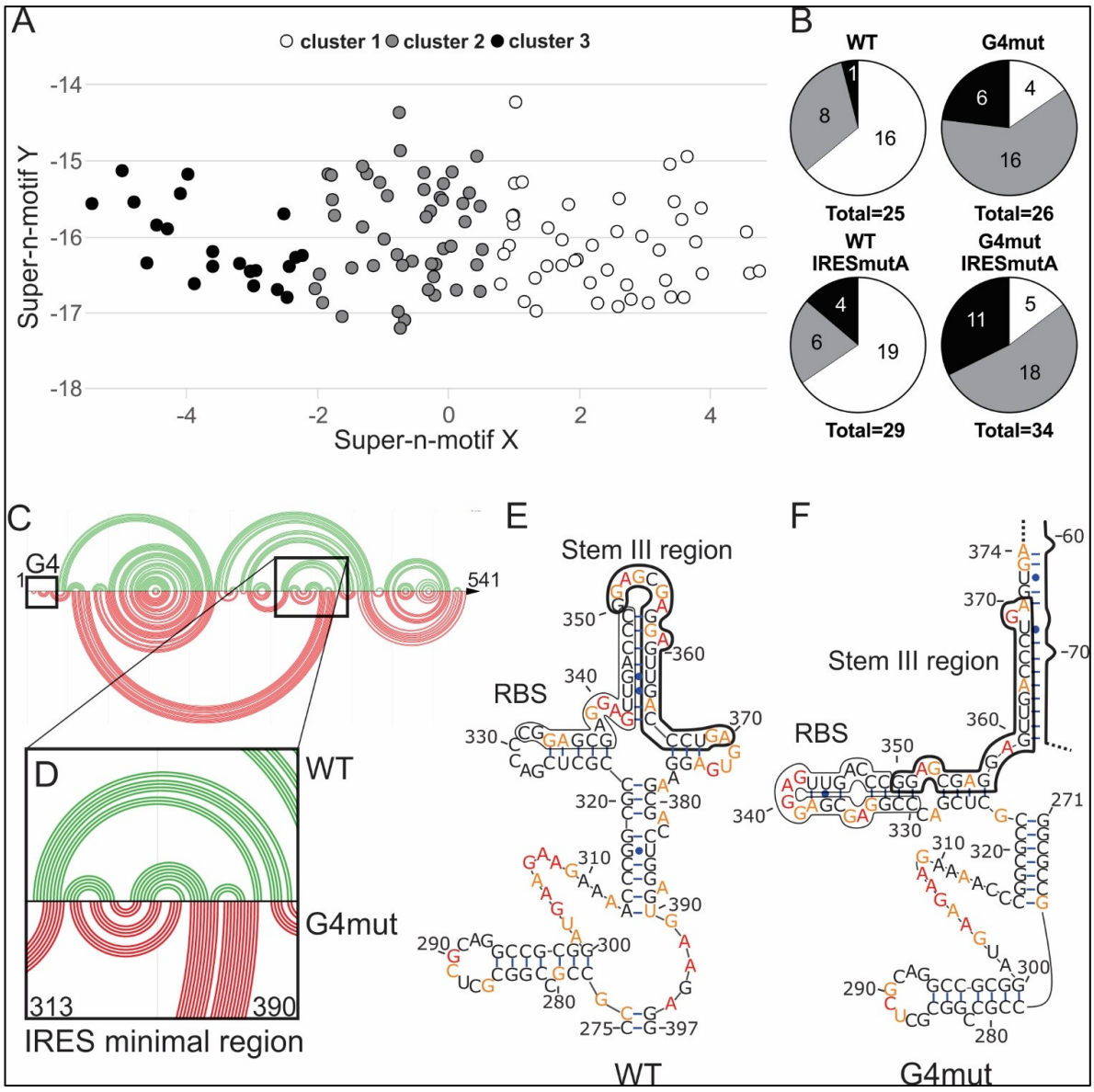



**Figure 42** – Effects of the rG4 and IRES mutations on the global secondary structure of the BAG 1 5'UTR, as analyzed by SHAPE.

(A) Super-n-motif representation of the 114 predicted secondary structures, separated into three clusters (white, cluster 1; gray, cluster 2; black, cluster 3). (B) Distribution in the three clusters of the predicted secondary structure of every sequences analyzed. (C) Arc-plot representation of the most stable predicted secondary structure of the complete WT (G4ss) sequence (green) compared to the rG4mut sequence (red). The rG4 region is boxed. (D) Close up of the arc-plot secondary structure of the IRES minimal region from nucleotide positions 313 to 390 (E, F). Most stable secondary structure of the minimal IRES region of the (E) WT (G4ss) sequence, and (F) rG4mut sequence. The color of the nucleotide represents its normalised SHAPE reactivity: black non-reactive; yellow, reactive; and, red, highly reactive. (*n*=2 for each of the 2 primers). Both the RBS and the Stem III region are boxed.

**The stability of the structural subdomains of the 5'UTR is affected by the rG4 formation**

To evaluate whether or not the rG4 folding affects the secondary structure surrounding either the start codons or the IRES subdomain of the 5'UTR, the most stable predicted secondary structures based on the SHAPE reactivity constraints for both the WT and the rG4mut sequences were compared in detail. The base-pairing, excluding the rG4 pairing, was represented using an arc-plot (**Figure 42C**). Of the 163 and 175 bp of the WT and rG4mut structures, respectively, 78 bp were identical. This represented 48% of the WT's and 45% of the rG4mut's total base-pairs, and they are shown as mirror images on the Arc-plot.

Stronger base-pairing and a higher energy of unfolding surrounding start codons are associated with less efficient translational initiation (Corley *et al.*, 2017). If the rG4 disruption resulted in the generation of more relaxed structural states for the start codons it could explain how protein synthesis is augmented in the rG4 mutant. However, no significant differences in the structures around the start codon regions could explain the change in the expression levels between the rG4 and the WT sequences, as all of the in-frame start codons, and even the uORF AUG-254 were in similarly accessible secondary structures. The base-pairing differences occur mostly in the middle region of the 5'UTR. Secondary structure representations of that region for each sequence (WT, rG4mut, WT-IRESmutA and rG4mut-IRESmutA) are presented in **Supplementary Figure S6 in Annexe 5**. Interestingly, this region (**Figure 42D**) corresponds to the previously characterised IRES structure (Pickering *et al.*, 2004). The secondary structure of the IRES region predicted here differ from that of the previous work mostly by a shift in the binding of the Stem III nucleotides, and by having globally more base-pairings (**Supplementary Figure S7 in Annexe 5**). However, the most flagrant alteration upon rG4 abolition is the long-range base-pairing of the nucleotides of the IRES regions, located at positions 360 to 380, with the nucleotides from positions 55 to 77 instead of the intrinsic folding observed for the WT sequence (**Figure 42E and F**, WT and rG4mut, respectively). This change results in a sliding offset in the base-pairs from the identified RBS and Stem III regions and affects the stability of both the previously defined Stem III region and the adjacent RBS. Evaluation of the changes, in terms of in minimum free energy (MFE) as measured by the RNAeval tool of the Vienna RNA package (Lorenz *et al.*, 2011), illustrated the differences in the predicted stabilities of these domains between

the various mutated sequences (**Table 9**). Globally, there is no difference in the stability of the complete 5'UTR secondary structure, with MFE ranging from -237.2 to -234.1 kcal/mol. However, the minimal IRES subdomain is more stable in the rG4mut folding (-105.7 kcal/mol) as compared to the WT (-74.0 kcal/mol). The disruption of the rG4 seems to shift the folding, making the IRES minimal region more stable. Based on the proposed mechanism of the IRES regulation of BAG-1 (Pickering *et al.*, 2004) a more "closed" structure might be more difficult to unfold and therefore impede the binding of the ITAF that is essential for the 40S ribosomal subunit's recruitment, and would thus explained the 20% decrease in IRES activity observed for the rG4mutant. An interesting perspective would be to measure the binding affinity of the ITAF depending on the global 5'UTR secondary structure.

**Table 9** Predicted minimum free energies (MFE) of the most stable secondary structures predicted by SHAPE for each mutant and region of the 5'UTR

| | **Minimum Free Energy (kcal/mol)** | | | |
|---|---|---|---|---|
| **Sequence region** | **WT** | **rG4mut** | **WT IRESmutA** | **rG4mut IRESmutA** |
| **Complete 5'UTR** | -236.07 | -234.06 | -236.07 | -237.20 |
| **Minimal IRES** | - 74.00 | -105.70 | -71.10 | -100.30 |
| **Stem-loop III** | -8.10 (11 bp) | -26.20 (16 bp) | -24.00 (13 bp) | -28.00 (15 bp) |
| **RBS** | -14.40 (13 bp) | -7.40 (7 bp) | -7.40 (7 bp) | -7.40 (7 bp) |

## CONCLUSION

Contrarily to the rG4s present in the VEGF and FGF-2 mRNAs, the BAG-1 rG4 is not in itself a structural part of the IRES domain. The BAG-1 rG4 instead possess a dual role. First of all, along with the repressive uORF, it acts as a roadblock that affects both the scanning and the translation of all in-frame isoforms in order to keep the overall BAG-1 protein synthesis at the right level under normal conditions. Secondly, it acts indirectly as a structural scaffold, with its presence allowing the maintenance of the global folding of the 5'UTR, as well as the folding of its internal subdomains such as the IRES secondary structure that is essential for translation under stress conditions. The BAG-1 rG4 is the first characterized rG4 with functions both in cap-dependent and independent translation.

The BAG-1 protein isoforms are anti-apoptotic proteins which were shown to be overexpressed in CRC and are associated with a poor prognosis. In this work, the post-transcriptional regulation of the BAG-1 mRNA was demonstrated in CRC cell lines and in paired healthy and tumoral tissues. The isoforms' translation levels could be repressed with specific ligands targeting the rG4 structure present at the 5'end of the BAG-1 mRNA's 5'UTR. This rG4 is located upstream of several *cis*-regulatory elements in the 5'UTR: alternative start codons, a non-canonical CUG start codon, an IRES and a repressive uORF newly identified in this work. Like other previously described rG4s located in 5'UTR, the BAG-1 rG4 represses the dominant cap-dependent translation of the three main protein isoforms. Nevertheless, the rG4's folding has a dual impact on translation, as it also affects the IRES-dependent translation even though it is not part of the IRES structure itself. Instead, the rG4 is responsible for the global maintenance of the 5'UTR's secondary structure. Its disruption by key G-to-A mutations triggered a shift in the secondary structure of the IRES subdomain located 300 nts away from the rG4 in the 5'UTR.

A proteogenomic analysis previously demonstrated that mRNA abundance is not a good predictor of protein abundance in colonic and rectal tumors (Zhang *et al.*, 2014). Recent studies have highlighted the alteration of translation regulation in various cancers (Robichaud *et al.*, 2018), including CRC (Provenzani *et al.*, 2006), in favor of alternative mechanisms, such as leaky scanning, re-initiation and IRES usage, that promote high proliferation, invasion and resistance to apoptosis and therapy (Sriram *et al.*, 2018). BAG-1 being a colorectal anti-apoptotic oncogene that is regulated by alternative translation mechanisms, represents a good model with which to study how the presence of an rG4, along with the different regulatory elements that are present in the 5'UTR, can affect protein synthesis. The translational repression of specific mRNAs by the use of small molecules targeting the rG4s located in the 5'UTR have been demonstrated (Miglietta *et al.*, 2017). Deciphering more examples like the rG4 of the BAG-1 mRNA 5'UTR could represent future avenues for therapies, as well as a better understanding of the mechanisms of action of rG4 on the translation regulation of other mRNAs possessing similar organisation of their 5'UTR.

## SUPPLEMENTARY DATA

**Annexe 5 :**
**Supplementary Material and Methods**
**Supplementary Figures and Legends S1-S7**
**Supplementary Tables S1-S5**

## ACKNOWLEDGEMENTS

The authors thank Jessica Gagné-Sansfaçon for sharing her expertise in the culture of the CRC cell lines. We also would like to thank Cameron Levins and Josiann Normandeau-Guimond for their technical assistance in both the cloning of the DNA bicistronic constructs and the SHAPE probing, Lubos Bauer for the 5'cap synthesis assays and Jean-Pierre Glouzon for the use of the StructurXplor software. R.J. was the recipient of a doctoral training scholarship from Fonds de Recherche du Québec Santé (FRQ-S). J-P.P. holds the Research Chair of the University of Sherbrooke in RNA Structure and Genomics. J-P.P., M.B. and N.R. are members of the Centre de Recherche du CHUS.

# DISCUSSION

L'objectif principal de cette thèse était de mieux comprendre l'impact du contexte nucléotidique dans le repliement et les fonctions des structures rG4 situées dans les 5'UTR des ARNm. Cependant, pour atteindre cet objectif plusieurs problématiques devaient être considérées. Entre autres, il était nécessaire d'avoir une méthode de détermination du repliement rG4 qui était informative, tout en permettant d'évaluer la structure secondaire du contexte nucléotide entourant la région rG4 présumée. Avec l'aide d'une meilleure méthode *in vitro* d'analyse, il est ainsi possible d'évaluer plusieurs séquences rG4 potentielles et de pouvoir mieux déterminer les paramètres permettant le repliement rG4. Le tout a permis d'améliorer la prédiction des rG4, mais surtout a permis de mieux identifier quels 5'UTR adoptent des rG4 et quels sont leurs impacts sur la régulation de l'expression des ARNm sur lesquels ils sont présents.

## Utilisation de la cartographie *in-line* sur des séquences variées

Le premier objectif spécifique de cette thèse était d'établir une méthode d'évaluation du repliement rG4 *in vitro* permettant d'étudier des séquences variées, dans des conditions plus représentatives du contexte biologique où ces structures se retrouvent. La première partie de cet objectif a été atteinte. Des séquences très variées ont pu être cartographiées grâce à la méthode *in-line* en comparant l'accessibilité des nucléotides en conditions $Li^+$ et $K^+$. À l'origine, des séquences correspondantes au motif canonique ont été testées (Article 1, Beaudoin *et al.*, 2013). Par la suite, la méthode a été utilisée pour étudier des séquences avec des motifs irréguliers tels que les rG4 avec de longues boucles centrales, pouvant aller jusqu'à 60 nt de long (Article 3, Jodoin *et al.*, 2014). Grâce à cette méthode, on en connaît aujourd'hui beaucoup plus sur l'étendue des séquences pouvant former un rG4. La méthode *in-line* a été utilisée par la suite par d'autres chercheurs pour tester un éventail encore plus grand de séquences irrégulières comme des rG4 situés dans les 5' et 3'UTR possédant de longues boucles 1 et 3 (Bolduc *et al.*, 2016). Des rG4 dans des séquences codantes ont aussi pu être évalués grâce à cette technique (Thandapani *et al.*, 2015). De plus, ce ne sont pas uniquement des séquences d'ARNm qui ont été évaluées, puisque la présence de rG4 dans

des ARN non codants comme les miARN et leurs précurseurs a aussi pu être étudiée grâce à cette méthode (Rouleau *et al.*, 2018) ainsi que les ARN guides utilisés dans la méthode CRISPR CAS9 (Moreno-Mateos *et al.*, 2015). Ceci démontre que la méthode *in-line* peut s'appliquer à des séquences rG4 potentielles de diversité d'origine dans le transcriptome et de positionnement dans un ARNm. Un point important, qui constitue un des atouts principaux de la méthode développée pour cette thèse est que toutes les séquences choisies étaient des séquences naturellement retrouvées dans le transcriptome et non des séquences artificielles, dessinées de toute pièce ou raccourcies à la région rG4 minimum, afin de pouvoir être étudiées comme cela est souvent nécessaire pour les autres méthodes *in vitro* comme le dichroïsme circulaire ou la RMN.

La méthode de cartographie *in-line* a donc permis de mesurer l'influence du contexte nucléotidique sur le repliement rG4. Dans l'ensemble, des séquences avec un contexte nucléotidique de 15 à 50 nt de part et d'autre des motifs rG4 prédits ont été utilisées. Cela représente des séquences d'une longueur maximale pouvant aller jusqu'à 131 nt et 149 nt dans le cadre des travaux présentés dans cette thèse. Si l'on considère un G4 minimal formé de l'empilement de deux tétrades reliées par trois boucles d'un seul nucléotide dont la longueur totale de la séquence est de 11 nt, ou encore un motif rG4 canonique $(G_3N_7)_3G_3$ qui fait 33 nt de long, cela signifie que la cartographie *in-line* peut s'appliquer pour mesurer un très large contexte extérieur au motif rG4. On peut même imaginer faire la cartographie de plusieurs courts motifs rG4 consécutifs.

En somme, avec une méthode permettant d'étudier des séquences aussi longues, on peut vraiment étudier une grande variété de facteurs en *cis* mentionnés en introduction qui affectent le repliement rG4 ; par exemple, les séquences adjacentes, ainsi que des séquences rG4 intrinsèques très variées (**Figure 7**). En ce sens, les travaux présentés dans les chapitres précédents ont pu montrer que les rG4 caractérisés par cartographie *in-line* peuvent être très hétérogènes. Principalement à l'Article 4 (Jodoin et Perreault, 2018), où les rG4 formés dans les ARNm associés au cancer colorectal possédaient un grand nombre de tétrades possibles, variées avec 2, 3 ou 4 G consécutifs, ainsi que des boucles composées de tous les nucléotides, incluant même des séries de G ou de C pouvant être très longues.

En bref, la méthode *in-line* a été éprouvée sur un large ensemble de séquences variées en termes de positionnement dans un transcrit et dans quelques ARN non codants, ayant un

motif rG4 régulier ou irrégulier et en différentes tailles du motif rG4 et du contexte flanquant. Cependant, ces études ont été limitées à des séquences issues du transcriptome humain. Cette méthode de cartographie serait tout aussi applicable pour des séquences de transcriptomes viral, bactérien ou d'autres eucaryotes. Ces transcriptomes pourraient avoir une composition différente en ratio G-C, et l'impact du contexte sur le repliement rG4 pourrait être différent.

Une limite de cette méthode est la longueur maximale de la séquence nucléotidique qui peut être résolue sur un gel de séquençage. Sur un gel de 10% polyacrylamide dénaturant, d'une épaisseur de 1 mm et d'une longueur de ~55 cm, avec un temps de migration de 2 h à 60-65 W, comme ce fut utilisé pour les résultats présentés ici, il est possible d'obtenir une séparation des bandes et une bonne résolution pour une séquence d'environ 150 nt. Ce faisant, les 10 à 12 premiers nucléotides ne sont pas facilement visibles à cause du front de migration causé par la présence de sels et le niveau de pureté de l'ARN à la suite de l'étape de précipitation. Afin d'obtenir une bonne séparation des bandes pour des séquences de longueurs supérieures à 150 nt, il faut répéter la migration en augmentant le temps de séparation à plus de 2 h. Les premiers nucléotides seront perdus, car ils « sortiront » du gel, mais les bandes correspondantes aux nucléotides situés aux positions 150 et plus seront mieux résolues. Si 2 gels avec 2 temps de migration sont effectués, il faudra ajouter une étape supplémentaire de normalisation et de quantification afin de combiner les intensités de clivage des 2 gels et déterminer les changements de structure secondaire pour la séquence complète. Une amélioration de la méthode *in-line* comme proposée pour l'étude des rG4 serait possible. Ces améliorations permettraient d'éviter les étapes plus laborieuses de la migration sur gel grâce au développement de l'électrophorèse capillaire, du marquage des acides nucléiques avec des fluorophores plutôt que par un phosphate radioactif, ainsi qu'aux avancées dans les réactions de transcription inverse. Des appareils de séquençage d'ADN sont utilisés aujourd'hui afin d'effectuer la cartographie *in-line* sur des structures secondaires canoniques (Weinrich *et al.*, 2018). L'avantage de l'électrophorèse capillaire est que des séquences beaucoup plus longues peuvent être analysées. De plus, la détection et la normalisation sont facilitées par l'analyse des pics de détection mesurés directement par l'appareil (Lee *et al.*, 2015). Il serait ainsi possible de marquer avec deux fluorophores différents la séquence ARN repliée en $Li^+$ et la même séquence repliée en conditions $K^+$ et de co-migrer en parallèle les ARN clivés pour ensuite directement mesurer la différence de

clivage au nucléotide près et évaluer le repliement rG4. La méthode de séparation des bandes et de quantification des gels présentée dans l'Article 1 a l'avantage d'être accessible pour des laboratoires de biochimie et de biologie moléculaire standards (Beaudoin *et al.*, 2013). Cette méthode n'utilise pas d'équipement spécialisé comme pour le dichroïsme circulaire ou la RMN, ni ne nécessite la synthèse de séquences avec des nucléotides modifiés chimiquement (déaza-GTP), comme la méthode *FOLDeR* qui sera décrite plus loin, qui entraîne des coûts élevés. C'est donc une méthode aussi informative qu'accessible. Les désavantages concernant l'utilisation de la radioactivité, les limites de la longueur des séquences analysables et la quantification pourront être contournés grâce aux avancées technologiques récentes mentionnées.

Une autre des limites de la cartographie *in-line* reste la détermination de la structure secondaire majoritaire adoptée. En effet, durant les 40 h de clivage de l'ARN, un mélange de structures secondaires peut être présent en solution, par exemple un équilibre entre une structure rG4 et une structure Watson-Crick alternative ou entre différentes conformations rG4 possibles lorsqu'il y a plus de 4 séries de G consécutives. Dans ce cas, le patron de clivage représentera la somme de toutes les conformations présentes. Si les différences de clivage entre la condition $Li^+$ et $K^+$ sont situées entre des séries de G, la formation rG4 « générale » sera confirmée, mais il ne sera pas possible de déterminer quelle conformation est dominante lorsque plusieurs sont possibles. On peut tenter d'observer les différentes conformations en effectuant des mutants ciblant seulement quelques séries de G pour « forcer » des conformations rG4 précises. Cette limite est cependant présente pour l'ensemble des techniques de cartographie de l'ARN en solution. La migration des ARN repliés sur gel natif pourrait permettre d'observer les différentes proportions des différentes conformations si celles-ci sont suffisamment différentes pour entraîner des changements de mobilité électrophorétique.

La détermination par cristallographie et par RMN de quelques structures rG4 permet aujourd'hui de faire des liens avec les « réactivités » observées des nucléotides à l'acétylation par la méthode SHAPE qui représente leur degré de flexibilité. En observant la structure rG4 de conformation parallèle de la séquence ARN TERRA obtenue par cristallographie (Collie *et al.*, 2010), on a pu voir que la transition du squelette phosphodiester du dernier G de chaque série vers la boucle courte rend le groupement 2'-OH très accessible et donc plus facilement

acétylable (Kwok *et al.*, 2016b). Il serait intéressant de comparer de même les patrons de clivage obtenus par la méthode *in-line* adaptée au rG4 avec les structures déterminées par cristal ou RMN afin de constater si l'accessibilité observée en condition $K^+$ concorde. Cela est d'autant plus intéressant qu'il a été observé que des nucléotides des boucles pouvaient former des interactions supplémentaires avec le squelette phosphodiester des G dans les tétrades, ou encore venir s'empiler sur les tétrades. Ces régions du squelette des résidus des boucles « stabilisées » par ces interactions supplémentaires pourraient être moins flexibles, ne pas former la conformation *in-line* et donc ne pas présenter de ratio supérieur $K^+/Li^+$. En général d'ailleurs, on observe que ce ne sont pas tous les nucléotides des boucles qui ont des ratios de clivage élevés. C'est pourquoi la combinaison de ces 2 méthodes pourrait être intéressante afin de confirmer des résultats obtenus séparément. On peut même imaginer qu'en comparant les patrons de clivage *in-line* de plusieurs structures déterminées par RMN, il sera même possible à l'avenir d'inférer une conformation rG4 précise à partir d'un patron de clivage ou selon l'intensité des pics d'électrophorèse capillaire.

### Cartographie dans des conditions *in vitro* plus représentatives du contexte biologique

La deuxième partie du premier objectif consistait à recourir à une méthode d'évaluation *in vitro* des rG4 dans des conditions plus représentatives du contexte biologique réel où ces structures se forment. Une de ces conditions est la concentration en ions dans la solution. Dans la méthode *in-line* proposée, les concentrations en potassium utilisées (100 à 150 mM) sont semblables aux concentrations biologiques. Les structures secondaires Watson-Crick sont quant à elles souvent plus sensibles à la présence de $Mg^{2+}$. Lors de l'incubation de 40 h, le $Mg^{2+}$ était présent à une concentration de 20 mM, ce qui est hautement supérieur à la concentration intracellulaire de 1 mM. Le pH est aussi ajusté à 8. Cette concentration de $Mg^{2+}$ et ce pH, tous deux plus élevés, sont nécessaires afin de faciliter l'hydrolyse du squelette et ces conditions sont aussi utilisées pour la détermination par *in-line* de structures secondaires canoniques. Afin d'être le plus près des conditions biologiques, c'est plutôt lors de l'étape du repliement de l'ARN et du refroidissement lent (*slow-cool)* que les conditions se doivent d'être respectées. Dans les conditions présentées ici, le repliement est effectué à pH 7.5, avec 100 mM KCl, mais sans $Mg^{2+}$. Cela pourrait être ajusté plus finement, mais en présence

d'ARN, afin de contrôler à quel moment l'auto-coupure peut être présente (et donc éviter la dégradation trop rapide de l'ARN) il est préférable d'éviter l'ajout de $Mg^{2+}$ lorsque l'on chauffe la solution. De plus, une possibilité de modification de la technique afin de se rapprocher des conditions intracellulaires ainsi que d'accélérer le protocole serait d'effectuer l'incubation durant 20 h à 37°C plutôt que 40 h à température pièce (Weinrich *et al.*, 2018).

L'atout indéniable de la méthode *in-line* en comparaison avec les aux autres méthodes traditionnellement utilisées pour l'étude des rG4, comme le dichroïsme circulaire et la dénaturation thermique, est l'utilisation de concentrations d'ARN beaucoup plus faibles. Selon la méthode *in-line,* l'ARN est présent en trace, donc en concentration très faible (< 1 nM) plutôt qu'en concentration µM. Cette faible concentration favorise les structures intramoléculaires plutôt qu'intermoléculaires et donc simule une condition beaucoup plus semblable aux conditions intracellulaires. Par contre, comme expliqué en introduction, l'environnement cellulaire est très encombré. C'est une condition qui favorise la formation G4. Il serait possible, afin de mieux simuler ces conditions, d'effectuer le repliement de l'ARN ainsi que l'incubation pour la cartographie *in-line* en présence d'agents encombrants comme le PEG et autres osmolytes.

### Utilisation de la méthode *in-line* afin d'évaluer l'impact de facteurs *trans*

Un autre atout de la méthode *in-line* est que celle-ci est très versatile afin d'évaluer l'effet des facteurs en *trans* qui affectent le repliement rG4 **(Figure 7)**. Bien que ces résultats n'aient pas été publiés, la cartographie *in-line* du candidat rG4 du 5'UTR de l'ARNm BAG-1 a été effectuée en présence des ligands chimiques (c'est-à-dire Phen-DC3 et PDS). Cela a servi de contrôle permettant de mesurer la stabilisation ou non de la structure par les ligands. De plus, toujours pour le rG4 de BAG-1, la cartographie a aussi pu être effectuée en présence d'ASO. Les ASO étaient des séquences de différentes longueurs, avec des nucléotides modifiés 2'-O-methyl ou LNA (*locked nucleic acids*) qui ciblaient la région de la boucle centrale plus longue du rG4 de BAG-1. Ces essais de cartographie ont pu démontrer la liaison de ces courts oligonucléotides et la perte de la formation du rG4 en solution. Ces essais *in-line* en présence de ces agents artificiels en *trans* permettent d'imiter la présence de séquences complémentaires compétitives ou de petites molécules thérapeutiques qui pourraient être présentes ou utilisées pour traiter des cellules.

Par contre, afin d'être le plus près des conditions intracellulaires, il serait nécessaire d'évaluer le repliement rG4 en présence de protéines et de RBP qui peuvent les lier. Ceci n'a pas été évalué pour la méthode *in-line* adaptée à l'étude des rG4 présentée ici. Cela serait évidemment possible, puisque la méthode *in-line* « classique » permet de vérifier l'effet de la liaison de protéines sur la structure secondaire adoptée (Huang *et al.*, 2019). Afin de l'adapter pour les rG4, il suffirait encore une fois d'effectuer la même expérience en comparant la liaison de la protéine et le repliement en présence de $Li^+$ comparativement au $K^+$. Ceci est une perspective intéressante pour la méthode puisque récemment plusieurs nouvelles protéines ont été identifiées pour reconnaître des rG4 en 5'UTR. Des travaux récents de *pull-down* de protéines liant l'ARN BAG-1 suivi d'analyse de spectrométrie de masse, effectués par François Bolduc dans le laboratoire Perreault, ont permis d'identifier des partenaires protéiques du rG4 de BAG-1. Il sera possible d'utiliser la méthode de cartographie *in-line* pour confirmer la liaison de ces protéines ainsi que pour mesurer leur effet sur le repliement du rG4 et des autres structures secondaires adjacentes dans l'UTR.

### Complémentarité du *in-line* avec les autres méthodes *in vitro* d'études des rG4

Il n'y a aucune méthode unique d'évaluation *in vitro* des structures secondaires d'ARN incluant les rG4 qui soit parfaite. Il est préférable de combiner plusieurs méthodes. Dans les travaux présentés dans cette thèse, le *in-line* a été combiné à des essais de fluorescence en présence de NMM (Article 4 (Jodoin et Perreault, 2018)). Des séquences possédant des nucléotides avec des ratios $K^+/Li^+ > 2$ étaient associées à des valeurs de fluorescence élevées. Il a pu être observé que lorsque les patrons de clivage étaient imprécis ou plus difficiles à interpréter, car les patrons étaient non reproductibles ou les positions des bandes étaient difficiles à identifier, cela était souvent associé à des séquences d'ARN difficiles à transcrire *in vitro,* mais aussi à des intensités de fluorescence avec NMM beaucoup plus faibles ou intermédiaires. Des exemples de ces cas sont les candidats SMAD2 et TCF7L1 à l'Article 4 (Figure S1 de (Jodoin et Perreault, 2018)). Cela démontre que les conclusions d'absence de formation rG4 formulées par l'absence de patrons de clivage différents entre $Li^+$ et $K^+$ par *in-line* étaient récapitulées en NMM. Les séquences formées de plusieurs motifs répétés et très G-riches sont beaucoup plus difficiles à étudier *in vitro* et ne semblent pas adopter de

structure rG4, mais plutôt des structures double-brin extrêmement stables ou forment en solution des structures intermoléculaires.

La formation du rG4 de BAG-1 a aussi été évaluée par essai d'arrêt de la transcriptase inverse (RTS, *Reverse transcriptase stalling assay*) en présence ou non des ligands TmPyP4, Phen-DC3 et PDS. En bref, ces résultats non publiés ont permis de confirmer les résultats obtenus par *in-line*. Cette seconde méthode est effectuée dans des conditions en solution très similaires et permet d'identifier le « dernier » G en 3' impliqué dans le rG4. Cela complémente l'étude *in vitro* d'une façon différente : en montrant que le rG4 est suffisamment stable pour bloquer l'enzyme de transcription inverse différemment selon la présence ou non de ligand. Par contre, la méthode RTS est beaucoup moins informative sur la structure secondaire de l'ensemble de la séquence que peut l'être la méthode *in-line*. Comparativement à la cartographie *in-line* ou cela n'est pas possible, des essais RTS peuvent être effectués dans des lysats d'extraits cellulaires, par contre d'autres contrôles doivent être effectués afin de confirmer que les pauses observées sont vraiment dues à la présence d'un rG4 (Weldon *et al.*, 2016).

Une hypothèse qui n'a pas été explorée et qui mériterait de l'être concerne la stabilité des rG4. L'objectif serait de comparer les patrons de clivage obtenus par *in-line* avec des analyses de dichroïsme circulaire et de dénaturation thermique. Le but serait de déterminer si des patrons de clivage « forts » (avec des nucléotides aux ratios $K^+$/ $Li^+$ >>2) sont associés à des rG4 aux spectres de dichroïsme circulaire avec des pics positifs et négatifs à 264 et 245 nm plus prononcés ou à des valeurs de Tm plus élevées (donc à des rG4 plus stables).

Une autre méthode impliquant la cartographie en solution a été développée récemment pour l'étude des rG4. Celle-ci est intitulée *FOLDeR* (Weldon *et al.*, 2017a). Cette méthode est similaire en principe avec la méthode *in-line* proposée ici, mais utilise un clivage avec des RNases, T1, T2 et V1 plutôt que la propriété naturelle d'auto-hydrolyse de l'ARN. La majeure différence consiste à l'utilisation comme contrôle négatif d'une séquence d'ARN modifiée avec des 7-déaza-G plutôt que par des mutations G/A dans les séries de G. Cette modification chimique change un azote en carbone à la position N7 de la guanine. C'est une des positions qui est impliquée dans les paires de bases Hoogsteen essentielles à la formation de la tétrade. Cela empêche donc la formation de rG4 sans affecter la formation des paires de bases G-C ou G-U. Contrairement aux mutations G/A dans les séries de G, ce contrôle négatif

n'affecte donc pas les autres structures secondaires possibles alternatives au rG4. Ce contrôle avec déaza-G pourrait aussi être utilisé dans des essais de cartographie *in-line* et permettrait en effet d'éviter les interrogations à savoir si la mutation G/A n'affecterait pas un motif de G en série essentiel à la liaison de facteurs, la formation d'autres structures essentielles ou ne créerait pas de structure secondaire avec le A substitué. La limitation de cette technique est qu'il faut acheter des oligonucléotides synthétisés avec la modification déaza-G aux endroits voulus, ce qui est beaucoup plus coûteux que la transcription *in vitro*, surtout si l'on veut analyser de longues séquences avec un large contexte nucléotidique. Il est possible dans la transcription *in vitro* de changer le ratio des rGTP et déaza-GTP qui vont être insérés par la polymérase, cependant l'ajout de la mutation déaza-GTP sera aléatoire et l'on ne pourra pas déterminer à quelles positions les G modifiés seront présents. L'autre option est de modifier tous les G de la séquence par des déaza-G, mais cela peut diminuer grandement l'efficacité de transcription. L'avantage par contre est que par la suite on peut comparer le patron de clivage uniquement en condition K+ avec et sans modifications déaza-G et ainsi déterminer les structures secondaires présentes avec le rG4 et alternatives sans le rG4 formé. Un autre avantage de cette technique comparativement à la méthode *in-line* est que l'évaluation de la structure a pu être effectuée par clivage à la RNase H dans des extraits cellulaires, ce qui se rapproche énormément des conditions naturelles.

**La cartographie *in-line* adaptée au rG4 est très informative**

La force de la cartographie *in-line* comparativement aux autres techniques demeure qu'en une seule réaction d'auto-coupure de faible quantité d'ARN en $Li^+$ et en $K^+$, on peut déterminer la flexibilité et donc la structure secondaire de tous les nucléotides de la séquence, et ce pour de longues séquences. Cette méthode permet d'aller beaucoup plus loin dans l'analyse des séquences qui peuvent adopter un rG4. La méthode ne détermine pas uniquement si un rG4 est formé ou non. Elle permet d'identifier toutes les boucles possibles et toutes les séries de G possibles d'une séquence, avec leur différent niveau de réactivité et de protection. Cet avantage a été utilisé dans l'analyse des structures rG4 associées au cancer colorectal présentée à l'Article 4 de cette thèse (Jodoin et Perreault, 2018). En analysant les patrons de clivage de chaque séquence, représentés à la **Figure 34** (où chaque nucléotide avec un ratio $K^+/Li^+$ élevé est indiqué avec une étoile et les séries de G protégées du clivage

sont encadrées) on peut analyser plus en détail les caractéristiques spécifiques de chaque rG4. De plus, cela a permis d'identifier des similarités entre les rG4 présents en 5'UTR des transcrits associés à une voie de signalisation ou mécanisme commun. L'ensemble des rG4 possibles de la séquence sont considérés en se fiant uniquement aux données empiriques recueillies et sans déterminer arbitrairement un rG4 « principal ». Les séries de G considérées sont toutes celles comprises entre la première série de G en amont d'un nucléotide réactif et la dernière série de G en aval du dernier nucléotide réactif qui sont formées de 2 G minimalement et qui sont non réactives. Les boucles considérées sont formées de tous les nucléotides situés entre ces séries de G déterminées.

Les caractéristiques spécifiques des séries de G et des boucles des différents rG4 associés aux trois voies de signalisation ont pu être comparées et les résultats sont résumés à la **Figure 43.**

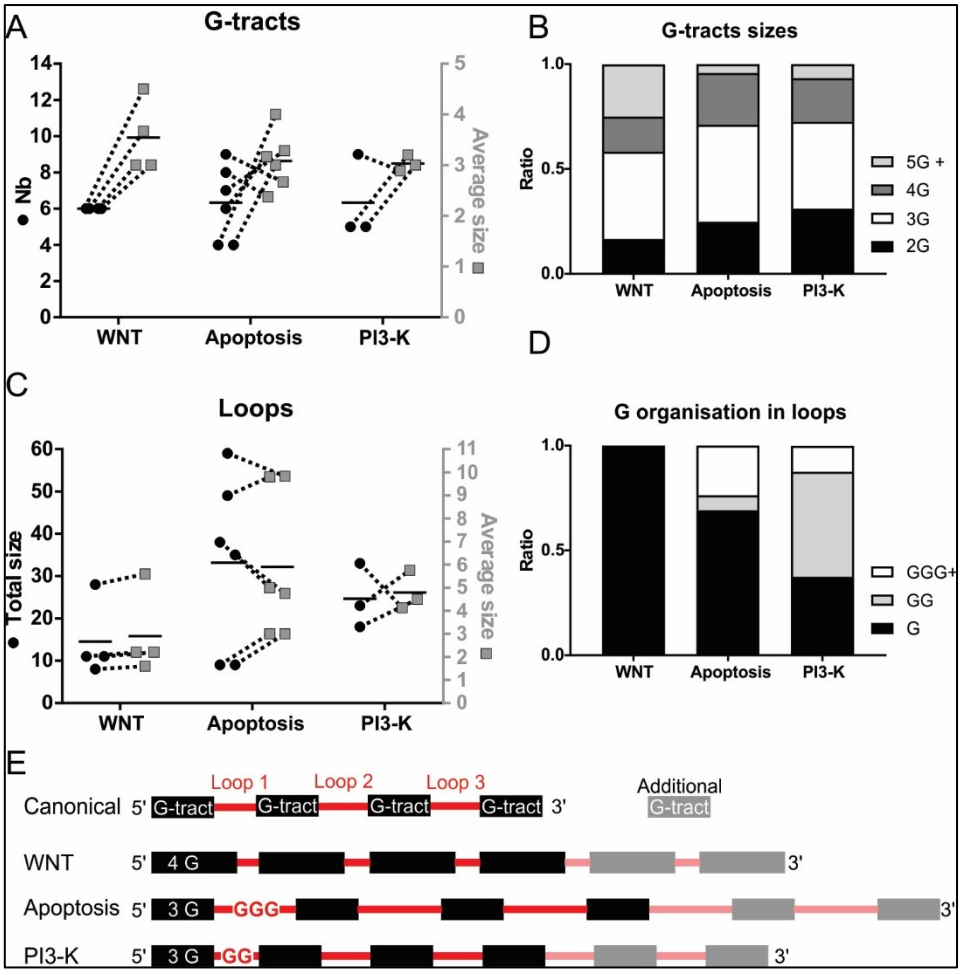



**Figure 43** – Comparaison des caractéristiques des rG4 situés en 5'UTR des ARNm associés aux voies de signalisation WNT, Apoptose ou PI3-K.

A) Séries de G des rG4 : Nombre de séries de G dans la séquence (axe *Y* gauche, cercle noir), moyenne de la longueur des séries de G (axe *Y* droit, carré gris). Pour les 2 caractéristiques, la moyenne est représentée par la courte ligne horizontale. Une paire composée d'un cercle et d'un carré relié par une ligne pointillée représente les caractéristiques d'un candidat rG4. B) Ratios du nombre de séries de G de différentes longueurs (2G à ≥ 5G) sur le nombre total de séries de G par candidat rG4. La moyenne des ratios pour chaque voie est présentée. C) Caractéristiques spécifiques des boucles : nombre total de nucléotides dans les boucles (axe *Y* gauche, cercle noir). Taille moyenne d'une boucle (axe *Y* droit, carré gris). Les moyennes sont représentées par les courtes lignes horizontales. Une paire composée d'un cercle et d'un carré relié par une ligne pointillée représente les caractéristiques d'un candidat rG4. D) Organisation des guanines (G) dans les boucles. Ratios du nombre de boucles possédant au moins 1 G, 2 G consécutifs (GG) ou 3 G consécutifs ou plus (GGG+) sur le nombre total de boucles avec G par candidat. La moyenne des ratios pour chaque voie est présentée. E) Schéma résumé des caractéristiques spécifiques des rG4 associés à chacune des trois voies. Les extrémités 5' et 3' du brin d'ARN sont indiquées. Les séries de G de différentes longueurs sont représentées par des rectangles de différentes longueurs. Les rectangles noirs représentent les 4 séries de G essentielles et les rectangles gris les séries de G supplémentaires. Les boucles sont représentées par des lignes rouges, les lignes plus longues représentent des boucles contenant plus de nucléotides. Les G en rouge représentent l'organisation générale des G présents dans les boucles (non limité à la boucle 1).

Cette analyse a permis de constater certaines similitudes entre tous les rG4 étudiés. Par exemple, la composition des nucléotides dans les boucles était identique pour les 3 voies (données non présentées). Toutefois, on peut observer quelques caractéristiques qui semblent particulières aux rG4 associés à certaines voies. Par exemple, les rG4 présents dans les 5'UTR des ARNm associés à la voie WNT possèdent plus de séries de G consécutifs plus longues de 4G ou plus que les candidats rG4 des 2 autres voies (**Figure 43B**). La taille des boucles varie aussi avec des candidats rG4 de la voie de l'Apoptose qui ont des boucles de 6-10 nt de long en moyenne comparativement à 3-6 nt pour les rG4 des 2 autres voies (**Figure 43C**). Les rG4 associés à l'apoptose ont aussi des boucles formées de séries de 3G réactives comparativement à la voie PI3-K qui montre des boucles avec 2G consécutifs réactifs (**Figure 43D**). On observe aussi que la région prône à la formation de rG4 est beaucoup plus étendue (à cause des boucles plus longues) pour les rG4 associés à l'Apoptose (**Figure 43E**).

Ces résultats préliminaires permettent d'émettre l'hypothèse qu'il pourrait exister des « sous-classes » de rG4 définies selon leurs caractéristiques spécifiques du nombre et de la longueur des séries de G, ainsi que de la taille et de la composition des boucles. Ces classes de motifs structuraux rG4 variés pourraient être reconnues par des RBP spécifiques à chacune de ces classes et pourraient indiquer un mode de corégulation de l'expression des transcrits impliqués dans des voies cellulaires ou métaboliques communes. Bien que pour le moment il n'y ait que peu d'études sur la détermination des motifs de reconnaissance spécifiques des protéines qui lient les G4 ADN ou ARN, des protéines ont des affinités connues pour certains éléments des G4. Par exemple, la liaison de la nucleoline est favorisée sur des G4 ADN à longues boucles (Lago *et al.*, 2017). La protéine hnRNP A1 quant à elle lie les rG4 en reconnaissant les bases azotées dans les boucles (Liu et Xu, 2018). Les seules structures rG4-RBP élucidées avec haute résolution par RMN et par cristallographie sont celles de FMRP lié à l'ARN *sc1*, une séquence G-riche sélectionnée *in vitro*. Le motif de reconnaissance de la protéine est l'interface rG4-duplex de la structure secondaire de *sc1* (Phan *et al.*, 2011 ; Vasilyev *et al.*, 2015). Récemment, l'interaction entre l'hélicase DHX36 et le G4 ADN du promoteur de c-MYC a aussi été élucidé. DHX36 reconnaît la tétrade supérieure du G4 d'ADN parallèle et le squelette phosphate de l'extrémité simple-brin en 3' afin de déplier le G4 un nucléotide à la fois (Chen *et al.*, 2018a). Des études bio-informatiques sont également

faites afin de prédire le motif de reconnaissance des G4 selon la composition en acide aminé des protéines (Brázda *et al.*, 2018). Bien sûr, avant de tirer la conclusion que des RBP corégulent un ensemble de rG4 spécifiques, les partenaires protéiques devront être mieux identifiés et caractérisés ainsi que plusieurs autres structures rG4 devront être déterminées et comparées. La technique de cartographie *in-line* adaptée aux rG4 pourra le permettre. Les travaux présentés à l'Article 4 sont parmi les seuls publiés ou des rG4 formés dans de longues séquences naturelles impliquées dans des voies communes sont comparés structurellement. En général, les travaux actuels se contentent d'identifier le rG4 le plus probable, souvent celui le plus semblable au motif canonique.

Un autre point important que l'étude systématique du repliement par cartographie *in-line* de plusieurs rG4 comme présentée dans l'Article 4 est que l'on constate qu'il y a rarement un seul rG4 possible dans la séquence. En général, il y a plus que 4 séries de G protégées et plus que trois boucles possibles. Parfois, comme dans le cas du candidat MAPK3, cela signifie que 2 rG4 consécutifs peuvent se replier. Cependant, pour la majorité des candidats étudiés, le nombre de séries de G est supérieur à 4, mais inférieur à 8. Cela signifie donc que les diverses combinaisons possibles de séries de G et de boucles formant les rG4 sont mutuellement exclusives. Tel qu'observé pour le candidat rG4 HIRA dans l'Article 3 (Jodoin *et al.*, 2014), abolir l'équilibre entre plusieurs conformations rG4 en insérant des mutations dans quelques séries de G, peut affecter les niveaux d'expression en cellule (**Figure 32**). En effet, la séquence naturelle du rG4 d'HIRA GGGCGGGCGGCGGCCGGAGGGCGGG, qui contient 7 séries de G pouvant former 35 combinaisons différentes de rG4 à 2 tétrades, inhibe plus fortement l'expression du gène rapporteur luciférase que la séquence mutée GGGCGGGC*AA*C*AA*CC*AA*AGGGCGGG (les A italiques correspondent aux mutations) qui formerait quant à elle un seul rG4 considéré plus stable avec 3 tétrades empilées. En conclusion, en plus de considérer la régulation possible formée entre le changement de l'équilibre entre la formation d'une structure rG4 et une structure secondaire Watson-Crick alternative, il faut considérer l'équilibre entre ces diverses conformations rG4 qui peut être tout aussi important ultimement pour réguler l'effet biologique de la structure secondaire. Les études plus détaillées des structures rG4 possibles grâce à la méthode *in-line* pourront permettre de mieux identifier quelles sont ces

confirmations plus spécifiques qui peuvent être reconnues, liées ou stabilisées par des protéines ou des ligands chimiques.

**Méthodes d'étude de la formation rG4 *in cellulo***

Évidemment, l'élucidation des structures secondaires adoptées *in cellulo* reste l'idéal afin de mieux comprendre le rôle biologique des rG4. Plusieurs techniques récentes ont été utilisées pour mesurer le repliement rG4 en conditions cellulaires.

La première technique, *rG4-seq* a consisté à isoler tous les ARN poly-adénylés de cellules HeLa et à effectuer la transcription inverse dans le lysat cellulaire en présence de conditions favorables aux rG4 (en présence de $K^+$ ou $K^+$ et ligand PDS) et en conditions défavorables (en présence de $Li^+$), pour ensuite effectuer le séquençage à haut débit. Puisque les rG4 stoppent la RT, des arrêts ou des pertes de qualité de séquençage en $K^+$ absents en $Li^+$ indiquent la présence d'une région rG4. Avec cette technique, 3 845 régions rG4 ont été identifiées en condition $K^+$ et 13 423 en présence de $K^+$ et du ligand PDS (Kwok *et al.*, 2016a).

Un second groupe a utilisé la technique de cartographie par DMS en cellule suivie de transcription inverse effectuée *in vitro* et *in cellulo* avec des cellules mESC, HEK293T et HeLa. L'idée étant que si les G sont impliqués dans des rG4 *in cellulo*, ils seront protégés de la modification par le DMS et reformeront après l'extraction des rG4 qui bloqueront la RT, alors que les G accessibles modifiés *in cellulo* par le DMS ne pourront pas reformer de rG4 et la RT sera complétée. Les conclusions de cette étude ont démontré que de très nombreuses régions rG4 sont formées *in vitro*, mais que la majorité est non repliée *in cellulo*. Ces résultats et certains aspects techniques du protocole sont fortement débattus (Kwok *et al.*, 2018), mais semblent renforcer l'idée que la formation de rG4 est dynamique, et que leur formation est régulée de façon transitoire par des RBP dans des processus précis. Finalement, un troisième groupe a tenté d'identifier ces rG4 « transitoires » en utilisant un nouveau type de ligand biotinylé spécifique aux rG4 permettant d'aller pêcher les ARN formant la structure et de les séquencer. La technique s'intitule *G4-RNA-specific precipitation with sequencing* (*G4RP-Seq*)(Yang *et al.*, 2018). Ils ont identifié plus de 300 gènes hautement exprimés avec des rG4 dans des cellules MCF-7. Par contre, cette technique ne permet pas d'identifier exactement où dans une séquence se situe le rG4.

Bien que ces études aient été effectuées dans des lignées cellulaires différentes, avec ou sans présence de ligands spécifiques au rG4, et qu'elles présentent plusieurs nuances entre elles, il reste intéressant de vérifier si les rG4 identifiés dans les travaux de cette thèse le sont aussi dans ces essais *in cellulo.* Si l'on considère que la formation des rG4 est régulée et s'effectue de façon transitoire, on ne s'attend pas à ce que les rG4 identifiés de façon *in vitro* ici le soit nécessairement de façons *in cellulo* dans ces études. De plus, les niveaux d'expression des transcrits peuvent varier entre les lignées utilisées et affecter les niveaux possibles de détection. En somme, le rG4 du 5'UTR de BAG-1 est identifié avec *rG4-seq*, ainsi qu'avec le *DMS-seq,* mais uniquement dans les cellules HEK293T, et est identifié dans les rG4 transitoires de *G4RP-seq*, malgré qu'il ne soit pas classé dans les gènes abondants. Les autres candidats rG4 identifiés dans cette thèse sont aussi différemment détectés ou non selon les trois méthodes. La détermination de la structure adoptée *in cellulo,* ainsi que la détermination des moments et des conditions cellulaires précis dans lesquelles les rG4 sont repliés constituent les futurs défis de l'étude des rG4.

**Prédiction des rG4**

Le second objectif spécifique de cette thèse était de développer un meilleur outil de prédiction des G4 d'ARN basé sur des facteurs pouvant affecter leur repliement autres que la simple présence du motif canonique. La popularité de l'étude des rG4 dans les dernières années, incluant les travaux présentés à l'Article 2 de cette thèse ont permis de mieux comprendre les paramètres influençant la formation de rG4 ((Beaudoin *et al.*, 2014). De très nombreux candidats rG4 dans des ARNm ont été analysés individuellement en présence de leur contexte nucléotidique adjacent dans les Articles 2, 3, et 4 présentés dans cette thèse, ainsi que par d'autres groupes de recherche (dont plusieurs sont résumés au **Tableau 2** de l'Introduction). Les analyses à haut débit *in cellulo* présentées ci-haut ont aussi augmenté le nombre de candidats rG4 possibles connus. À ce jour, plusieurs banques de données de séquences adoptant la formation G4 existent et sont répertoriées dans le **Tableau A2 en Annexe 6** de cette thèse. Il reste tout de même frappant qu'il n'existe que 2 bases de données sur 11 qui sont spécifiques aux G4 d'ARN.

L'identification des nouveaux rG4 a permis de confirmer que ceux-ci sont en effet très diversifiés, possédant des caractéristiques considérées comme atypiques telles que de

longues boucles et des renflements. En effet, la méthode *rG4-seq* en présence de $K^+$ a identifié que seulement 30% des régions rG4 identifiées correspondaient au motif canonique et donc la majorité, 70%, étaient « irréguliers » avec seulement 2 tétrades, avec la présence de renflements ou de longues boucles (Kwok *et al.*, 2016a). Cela confirme que la prédiction basée sur la recherche de motif doit évoluer afin d'être applicable aux rG4. C'est un défi auquel plusieurs chercheurs se sont attaqués et il existe dorénavant plusieurs outils bio-informatiques qui permettent de prédire des rG4 avec de longues boucles, des renflements ainsi que des mésappariements et ceux-ci sont présentés dans le **Tableau A1 en Annexe 1**.

**Impact des séquences adjacentes dans le repliement des G4**

L'hypothèse que la présence de C consécutifs dans l'environnement proximal du motif G4 était néfaste au repliement G4, ainsi que la possibilité d'utiliser la composition nucléotidique du contexte afin de mieux prédire les G4 qui a été utilisée comme prémisse pour le développement du score cG/cC présenté ici, a aussi été développée pour les G4 d'ADN. Le score développé s'intitule G4H pour *G4Hunter*. Dans ce système, les nucléotides A et T se font attribuer un score de 0, les G seuls un score de 1, les G dans des doublets un score de 2, les G dans des triplets un score de 3 et ainsi de suite. Le score est identique, mais sous forme négative pour les C (un seul C a un score de -1, dans une suite CC chaque C a un score attribué de -2, etc.). Le score G4H est donc calculé comme étant la moyenne arithmétique des valeurs de chaque nucléotide dans une fenêtre de taille définie. La limite de détection a été statuée à 0,9. Il faut donc que la moyenne de la fenêtre soit supérieure à ce nombre pour que le contexte soit favorable à la formation de G4.

Le score G4H est très semblable au score cG/cC dans l'optique qu'il permet aussi de quantifier à l'aide d'une métrique le contexte compétitif Watson-Crick, mais il a été testé pour des séquences d'ADN. La force de ce travail comparativement au score cG/cC est que le G4H a été testé sur un nombre beaucoup plus grand de séquences au départ, c'est-à-dire une banque de données testées expérimentalement de 392 séquences (298 G4 et 94 non-G4, négatif). Cela était bien sûr facilité par l'abondance de travaux sur les G4 d'ADN comparativement à ceux d'ARN à la même époque. Tel qu'observé à l'Article 4 de cette thèse, lorsqu'on souhaite prédire des G4 d'ARN, le score cG/cC reste le plus sensible et le plus spécifique comparativement au score G4H.

Le score cG/cC développé reste une simple métrique, basé sur un calcul mathématique intuitif et additif ou « plus de G en série sont associés à une plus grande probabilité de G4 et moins de C en série résulte en moins de compétition ». Afin de le simplifier, le score aurait pu être basé sur l'élaboration d'un facteur multiplicatif plutôt que par des sommes arbitrairement choisies de 10, 20 ou 30 pour des G seuls, en série GG et en série GGG. Cependant, même dans sa formule initiale quasi naïve, c'est une métrique supplémentaire qui semble efficace lorsqu'ajoutée à la recherche de motifs tel que démontré aux Articles 2 et 4 (Beaudoin *et al.*, 2014 ; Jodoin et Perreault, 2018). Cela complémente la prédiction et permet de filtrer les faux positifs. Par contre, dans la façon présentée ici, le score cG/cC n'est pas utilisé pour interroger le transcriptome, il est utilisé uniquement après avoir identifié des motifs PG4 avec les limites que cela implique sur la diversité des rG4 prédits. L'utilisation du score cG/cC pour la prédiction *de novo* de rG4 au travers du transcriptome et ainsi éviter la recherche de motif prédéterminé est possible. Cela s'effectue en calculant le score dans plusieurs fenêtres défilantes. Cette utilisation du score a été étendue dans les travaux de mes collègues Jean-Michel Garant (Garant *et al.*, 2015, 2017, 2018) et Sarah Belhamiti (manuscrit en préparation). De plus, le score cG/cC de n'importe quelle séquence d'ARN peut maintenant être facilement calculé grâce à l'outil web *G4RNAscreener* (URL http://scottgroup.med.usherbrooke.ca/G4RNA_screener/) ainsi qu'avec l'outil web *Putative G-quadruplex prediction tool* de la banque de données *G4IPDB* (URL http://bsbe.iiti.ac.in/bsbe/ipdb/pattern2.php). Cela démontre l'acceptation de ce score par la communauté scientifique étudiant les G4, en espérant que son utilisation se répande à l'ensemble des scientifiques intéressés à l'impact des structures secondaires d'ARN sur la régulation de l'expression.

**Prédiction des rG4 par apprentissage automatisé**

L'évolution du domaine de la prédiction des séquences G4 se fera grâce à l'utilisation de plus en plus répandue des différents systèmes d'apprentissage automatisé (*machine-learning)*. Les travaux les plus récents tentent maintenant d'inclure les imperfections et les caractéristiques non canoniques des G4 dans les prédictions avec les outils *pqsfinder* et *impG4finder* (Hon *et al.*, 2017 ; Varizhuk *et al.*, 2014). Bien entendu, pour que ces outils bio-informatiques soient efficaces les modèles doivent être entraînés sur des banques de données larges et bien

diversifiées. Suite à la publication des travaux sur le score cG/cC en 2014, Article 2 (Beaudoin *et al.*, 2014), plusieurs améliorations dans la prédiction des rG4 ont été apportées. Un nouveau score intitulé G4NN a été développé grâce à l'utilisation de la toute première banque de donnée de rG4 validés expérimentalement ainsi qu'avec leurs séquences contrôles négatives associées, c'est-à-dire la banque de donnée G4RNA (Garant *et al.*, 2015). Celle-ci a été utilisée afin d'entraîner un réseau de neurones artificiels à prédire des rG4 basés sur l'homologie avec les rG4 caractérisés de la banque comparativement à des séquences aléatoires du transcriptome. Cette méthode n'est pas biaisée par des assomptions sur la présence d'un motif, de taille prédéfinie des boucles ou de la présence d'un contexte nucléotidique adjacent, puisque le réseau de neurones peut considérer plusieurs autres aspects comme le ratio dinucléotidique qui peut considérer tous les nucléotides et non uniquement les G et les C par exemple (Garant *et al.*, 2017). Par contre, cette banque de données est limitée aux séquences rG4 qui ont été étudiées par la communauté scientifique et donc contient principalement les rG4 « canoniques » qui ont été généralement prédits par la présence du motif et non pas un ensemble varié.

Tel que mentionné, il existe maintenant plusieurs banques de données sur les G4, mais la majorité sont des banques de données de G4 prédits (sur la base du motif canonique), très peu sont basées sur des G4 démontrés expérimentalement dans des conditions physiologiques ou qui considèrent le contexte flanquant. De plus, de ce nombre, seulement deux sont des banques de données concernant les rG4 : G4RNA déjà mentionné et la liste des rG4 identifiés par un seul groupe par la technique *rG4-Seq* (**Tableau A1 Annexe 1**). Le succès futur de la prédiction des rG4 sera en fonction du développement de grandes banques de données, basées sur des données expérimentales rigoureuses.

Tout récemment, l'outil de prédiction *Quadron* a été présenté (Sahakyan *et al.*, 2017). C'est un modèle de prédiction par apprentissage automatisé qui a été entraîné sur la très large base de données des G4 d'ADN identifiés dans l'ensemble du génome grâce à la technique *G4-Seq*. Avec cette technique, les G4 sont identifiés grâce à un pourcentage élevé de mésappariements lors du séquençage (mesuré en pourcentage de mésappariements, *mismatch,* mm%), mais cela n'indique pas exactement les frontières de la région G4. Pour contourner ce problème, ils ont cherché avec le motif G4 consensus étendu (c'est-à-dire des séries de 3G séparées par des boucles jusqu'à 12 nt de long) les régions PG4 qui avaient un

mm% élevé et considéré en plus des régions flanquantes de 50 nt de chaque côté. L'utilisation d'une séquence PG4 entourée de 50 nt de part et d'autre est identique à ce qui a été fait lors des travaux concernant le score cG/cC. Par la suite, ils ont entraîné un système d'apprentissage automatisé pour identifier et considérer avec différents poids des éléments autant de la séquence primaire du PG4 que du contexte nucléotidique qui influenceraient la formation ou non des G4. Leur modèle final de prédiction permet de considérer 119 éléments de la structure primaire, et la validation a montré d'excellents taux de prédiction. De cette analyse complexe, ils ont pu ressortir les éléments utilisés dans la prédiction qui sont les plus favorables à la formation du G4. Ces éléments sont la présence de longues séries de G dans la région PG4 et surtout la présence de séries de 2G et plus dans le contexte flanquant. Les éléments les plus défavorables à la formation de G4 sont des boucles plus longues et la présence de séries de C dans le contexte flanquant.

En somme, ce système extrêmement sophistiqué d'analyse sur l'ensemble du génome à l'aide des outils bio-informatiques les plus récents confirme ce qui avait été déduit instinctivement lors de la conception du score cG/cC, soit que la présence de G consécutifs et de C consécutifs dans le contexte flanquant des régions PG4 influencent grandement la formation du G4 et qu'il est nécessaire de les considérer pour améliorer les prédictions.

**Fonctions biologiques des rG4**

Le troisième objectif spécifique de cette thèse était de déterminer si les G4 d'ARN en 5'UTR sont enrichis dans des voies biologiques particulières et par quels mécanismes ils affectent l'expression des ARNm sur lesquels ils se retrouvent. La première partie de cet objectif a été atteinte principalement dans l'Article 4 (Jodoin et Perreault, 2018). Grâce à une analyse d'ontologie, le cancer colorectal a été identifié comme une condition biologique où plusieurs ARNm qui y sont associés possèdent des rG4 en 5'UTR. Le repliement rG4 *in vitro* des candidats de cette liste a ensuite été caractérisé plus en détail. Néanmoins, il y avait six autres voies biologiques identifiées pour lesquelles nous n'avons pas poursuivi l'analyse qui sont aussi d'intérêt. Les études de prédictions et d'analyse sur l'ensemble du génome démontrent que les rG4 sont enrichis dans les oncogènes (Eddy et Maizels, 2006 ; Lung Chan *et al.*, 2018). Donc, il n'est pas surprenant que des voies enrichies, en plus du cancer colorectal, 2 autres soient aussi des cancers, plus particulièrement des leucémies. L'analyse d'ontologie a été effectuée en utilisant uniquement une courte liste de candidats rG4 potentiels respectant

le motif canonique et une variation avec des boucles plus longues. Maintenant que plusieurs nouveaux outils de prédiction de rG4 et de plus grandes banques de données existent, il pourrait être intéressant de répéter ces analyses d'enrichissement.

L'analyse des structures rG4 en 5'UTR d'ARNm impliqués dans les voies de signalisation WNT, de l'apoptose et de PI3-K, dérégulées dans le cancer colorectal, combinée à l'effet des rG4 sur la traduction suggère que les rG4 pourraient être impliqués dans la corégulation de la traduction. Des études ont démontré l'importance de la régulation traductionnelle dans la carcinogenèse en général (Robichaud *et al.*, 2018) et spécifiquement pour le cancer colorectal également (Provenzani *et al.*, 2006 ; Zhang *et al.*, 2014). De plus, des RBP sont surexprimées dans les tumeurs colorectales, telles que Lin28A et B, MSI1 et 2, IGF2BP, HuR, CELF1 et RBM3 (Chatterji et Rustgi, 2018 ; Mukohyama *et al.*, 2017). Cette dérégulation des RBP affecte l'initiation, la progression et le potentiel métastatique des tumeurs, entre autres en créant le profil d'expression des cellules souches cancéreuses (Mukohyama *et al.*, 2017). Certaines de ces RBP affectent la traduction de leurs ARNm cibles en liant les régions 5' ou 3'UTR (Chatterji et Rustgi, 2018). Des liens directs ou indirects entre la formation de rG4 et le cancer colorectal peuvent aussi être établis grâce à des études qui seront décrites ici.

Tout d'abord, un rG4 a été identifié dans le long ARN non codant GSEC (*G-quadruplex-forming sequence containing lncRNA*) (Matsumura *et al.*, 2017). Ce lncARN est surexprimé dans le cancer colorectal et serait impliqué dans la migration cellulaire. Ensuite, des travaux récents ont démontré que l'inhibition de la biosynthèse des polyamines permettrait de cibler le remodelage des canaux $Ca^{2+}$ dans des cellules du cancer du côlon (Gutiérrez *et al.*, 2019). Cette stratégie thérapeutique est efficace, car elle pourrait permettre aussi de réprimer l'enzyme CHSY1 (*Chondroitin synthase- 1)* qui stimule la prolifération et réprime l'apoptose des cellules colorectales cancéreuses (Zeng *et al.*, 2018). En effet, la synthèse protéique de CHSY1 est stimulée par la présence de polyamines. Les polyamines permettent de défaire une structure rG4 située dans le 5-UTR de l'ARNm CHSY1. En absence du rG4, son expression est fortement augmentée (Yamaguchi *et al.*, 2018). De plus, la voie de synthèse complète des polyamines semble être corégulée par la présence de rG4 formés de 2 tétrades dans les 5'UTR de plusieurs des ARNm codants pour les enzymes de

cette voie (Lightfoot *et al.*, 2018). Donc le lien entre la synthèse des polyamines et le cancer colorectal pourrait être médié par la formation de rG4.

Un autre aspect important de la carcinogenèse colorectale est la dérégulation des voies de signalisation de MYC et de WNT. Une voie thérapeutique prometteuse est de cibler la traduction de MYC (Wiegering *et al.*, 2015). Pour ce faire, on utilise un composé appelé silvestrol qui inhibe l'hélicase eIF4A du complexe d'initiation de la traduction (Cencic *et al.*, 2012). Plusieurs transcrits d'oncogènes sont dépendants à cette hélicase pour leur traduction. Les transcrits dont la traduction est affectée par le silvestrol ont souvent des 5'UTR plus longs et plus structurés que la moyenne. De plus, ils sont enrichis pour la présence d'un motif GGCGGCGGCGG caractéristique d'un rG4. La liaison spécifique de eIF4A aux rG4 ou à la séquence G-riche uniquement est cependant un sujet débattu (Waldron *et al.*, 2018 ; Wolfe *et al.*, 2014). Pour revenir plus spécifiquement sur la voie WNT, il a été démontré que le traitement au silvestrol de tumeurs mammaires affectait la traduction d'ARNm associés à cette voie et que ceux-ci possédaient aussi le motif GGC répété. La suractivation de la voie WNT entraîne la phosphorylation de PDCD4 et sa dégradation, créant ainsi la perte de l'inhibition de PDCD4 sur eIF4A. C'est l'activité augmentée de l'hélicase eIF4A qui pourrait expliquer la plus grande traduction des oncogènes ou des gènes associés à WNT qui ont des 5'UTR structurés et pouvant possiblement former des rG4. Il y a donc 2 lignes d'évidences qui semblent converger. D'abord, les rG4 en 5'UTR sont reconnus pour réprimer la traduction, et lorsque l'on cherche des voies enrichies pour la présence de rG4 dans les ARNm, des voies dérégulées dans le cancer sont identifiées. Inversement, en ciblant la traduction dans des cancers et spécifiquement une hélicase, les transcrits affectés possèdent des motifs probables de formation de rG4.

La dérégulation de la traduction est un phénomène connu de la carcinogenèse (Robichaud *et al.*, 2018), et comme décrit précédemment, plusieurs RBP sont surexprimées, entre autres dans les tumeurs colorectales. Cependant, on néglige souvent de considérer plus sérieusement comment la structure secondaire adoptée par les ARNm affecte leur traduction et leur liaison par des protéines, ou encore si la structure secondaire est modifiée dans un cancer. Souvent, un seul motif court commun (MEME) de 6 à 12 nt est identifié sur les ARNm traductionnellement dérégulés. De toutes les RBP et hélicases suggérées pour reconnaître et moduler les rG4, qui sont mentionnées dans le **Tableau 1**, l'étude systématique

des structures adoptées et des caractéristiques particulières des rG4 reconnus n'est pas faite. Il serait très pertinent de le faire pour confirmer l'hypothèse que les rG4 puissent former des sous-groupes avec des caractéristiques semblables et servir d'éléments de corégulation de la traduction, entre autres dans des voies reliées à la carcinogenèse.

Bien entendu, cela représente une perspective, le travail présenté à l'Article 4 a pu démontrer une association entre la présence de rG4 et une voie biologique et une confirmation *in vitro* des structures rG4 adoptées. Par contre, cela n'a pas permis d'en savoir plus sur le mécanisme d'action des rG4 dans ce contexte. Brièvement, l'effet des rG4 sur l'expression d'un gène rapporteur a été mesuré dans des lignées colorectales pour 3 candidats. Cependant, l'analyse n'a pas été poursuivie afin de déterminer si l'effet des rG4 était transcriptionnel ou traductionnel.

### Mécanismes d'action des rG4

Le second aspect du troisième objectif spécifique était d'évaluer le mécanisme d'action des rG4 situés en 5'UTR. Les candidats rG4 évalués *in cellulo* dans les Articles 3, 4 et 5 réprimaient l'expression d'un gène rapporteur luciférase suggérant un effet sur la traduction tel que décrit dans la littérature générale sur les rG4. Cependant, plus spécifiquement, c'est le candidat rG4 de BAG-1 dont l'effet sur la traduction a été décortiqué en plus amples détails afin de proposer un mécanisme d'action.

#### Présence d'un rG4 près de l'extrémité 5' du transcrit

Un des mécanismes d'action proposés expliquant l'effet répresseur des rG4 sur la traduction est la stabilité élevée de la structure causant un encombrement stérique qui est néfaste lorsqu'il survient trop près de l'extrémité 5' d'un transcrit. Cette logique est basée sur la littérature scientifique étendue à propos des structures secondaires canoniques très stables à l'extrémité 5' qui nuisent à la traduction (Babendure *et al.*, 2006 ; Sagliocco *et al.*, 1993). Le rG4 du transcrit BAG-1 est situé à 6 nt de l'extrémité 5' et donc son effet répresseur pourrait être expliqué par cette proximité. Par contre, les essais enzymatiques non publiés de liaison de eIF4E ou de synthèse de la coiffe en présence ou non du rG4, effectués par Lubos Bauer dans le laboratoire Perreault, ne démontrent pas d'effets répresseurs du rG4 sur ces aspects.

L'organisation particulière du 5'UTR de BAG-1 avec un élément structural très stable, le rG4, à l'extrémité 5', en amont d'un élément IRES offre des similarités avec les

éléments TIE (*translation inhibitory elements)* identifiés dans le transcriptome du poisson-zèbre. Ces éléments présents dans les 5'UTR des gènes HOX associés au développement sont de petites structures secondaires très stables qui répriment la traduction coiffe-dépendante de ces transcrits. Ces transcrits possèdent également dans leurs 5'UTR un élément IRES. Dans le processus du développement de l'embryon, la synthèse protéique des gènes HOX doit s'effectuer de façon extrêmement régulée et s'effectue alors par la traduction coiffe-indépendante par l'IRES (Xue *et al.*, 2015). Les 2 éléments : structure secondaire stable à l'extrémité 5' et IRES en aval sont essentiel pour le mécanisme de régulation. Le rG4 dans le transcrit BAG-1 pourrait jouer un rôle semblable à l'élément TIE.

**Présence de rG4 avec des codons d'initiations alternatifs**

Un deuxième élément particulier dans le 5'UTR de BAG-1 est la présence de plusieurs codons d'initiation dans le même cadre de lecture, ainsi que la présence d'un codon non canonique CUG (**Figure 35**). La mutagenèse dirigée de chacun de ces codons a permis de montrer que le rG4 situé en amont avait un effet répressif peu importe le codon d'initiation utilisé. Lorsque plusieurs codons d'initiations sont présents, deux critères influenceront le choix du codon pour l'initiation. Le premier critère est la présence du contexte nucléotidique l'entourant intitulé le contexte Kozak. Le contexte idéal est GCCACC**AUG**GG, mais plusieurs contextes sont possibles. Deux études ont systématiquement analysé l'efficacité d'initiation de tous les contextes nucléotidiques possibles entourant les codons d'initiation canonique AUG et toutes les variations non canoniques CUG, UUG et GUG (Diaz de Arce *et al.*, 2018 ; Noderer *et al.*, 2014). L'efficacité du contexte des codons d'initiation du 5'UTR de BAG-1 concorde avec les niveaux exprimés de chacun des isoformes protéiques. Le deuxième critère affectant le choix du codon d'initiation est son accessibilité en termes de structure secondaire (Corley *et al.*, 2017). Puisque la structure secondaire complète du 5'UTR de BAG-1 a été déterminée, l'accessibilité des codons d'initiations est connue (**Figure 42 et Figures 44 et 45 en Annexe 7**). Tous les codons d'initiations, même l'AUG-254 de l'uORF ont des réactivités SHAPE intermédiaires ou élevées. Ils sont donc simple-brins et accessibles. Cette accessibilité de change pas non plus suite à la mutation du rG4.

L'organisation du 5'UTR de BAG-1 avec des codons d'initiation alternatifs, un IRES, un uORF et un rG4 est semblable à l'organisation du transcrit de VEGFA dont chacun des éléments ont été individuellement évalués (Agrawal *et al.*, 2013 ; Arcondéguy *et al.*, 2013 ;

Bastide *et al.*, 2008 ; Cammas *et al.*, 2015 ; Morris *et al.*, 2010). Le rG4 du transcrit VEGFA semble aussi affecter la traduction. De plus, des résultats récents utilisant le *ribosome profiling* démontrent que plusieurs ARNm possèdent des codons de départ alternatifs ainsi que des extensions N-terminales (Ivanov *et al.*, 2011 ; Fritsch *et al.*, 2012 ; Ingolia *et al.*, 2011). Une revue rapide indique que la conjonction de la présence de rG4 et de sites d'initiation de la traduction alternatifs est fréquente. En effet, à partir d'un ensemble de 70 5'UTR d'ARNm avec des extensions N-terminales découlant de l'utilisation d'un codon d'initiation non canonique (Ivanov *et al.*, 2011), il y a 31 de ces UTR qui ont minimalement un rG4 prédit grâce à l'outil *G4Screener* en amont du codon d'initiation alternatif. Cette analyse démontre que la situation présente dans le 5'UTR de BAG-1 n'est pas unique, et qu'elle représente peut-être un mode de régulation de l'utilisation de codon de départ alternatif plus répandu.

**Présence de rG4 avec un uORF**

Les uORF sont de courtes séquences codantes situées en amont du cadre de lecture principal d'un transcrit. Généralement, ils sont des éléments répresseurs de la traduction qui lors de l'étape du *scanning* dévient la machinerie d'initiation de la traduction sur leur AUG limitant ainsi l'initiation au codon de départ du cadre de lecture principal du transcrit. L'ampleur de leurs effets répressifs varie selon leurs tailles, leurs distances par rapport à l'extrémité 5' et au codon de départ du cadre de lecture principal, leurs structures secondaires, ainsi que d'autres éléments en *cis* à proximité (Barbosa *et al.*, 2013 ; Johnstone *et al.*, 2016). L'impact d'une structure rG4 dans le contexte d'un uORF est un phénomène peu exploré.

Le 5'UTR de VEGFA, a beaucoup en commun avec BAG-1, avec la présence d'un rG4 et d'un IRES, mais aussi par la présence d'un uORF en 5'UTR (Bastide *et al.*, 2008). Contrairement à BAG-1, l'uORF est situé en 3' de l'IRES de VEGFA résultant en une organisation légèrement différente des éléments *cis* du 5'UTR. L'interaction entre le rG4 de VEGFA et l'uORF n'a pas été mesurée. Il y a aussi des similarités entre le mécanisme de régulation de la traduction du 5'UTR de BAG-1 avec le 5'UTR de l'ARNm CAT-1 (Yaman *et al.*, 2003). Dans le cas de CAT-1, la traduction de l'uORF situé en 5' entraîne le dépliement d'une structure secondaire inhibitrice qui permet d'activer une structure IRES située en 3'. Pour BAG-1, le repliement rG4 semble être important pour la structure globale et sa perte entraîne des changements conformationnels qui affectent l'efficacité de l'IRES. Dans le

5’UTR de BAG-1, le rG4 et le uORF semblent être 2 éléments répresseurs distincts de la traduction coiffe-dépendante, mais il reste que la question n’y a pas été répondue à savoir si le rG4 réprime aussi la traduction de l’uORF. Puisque le rG4 nuit à la traduction de tous les isoformes de BAG-1 qui sont situés en aval, il pourrait donc également nuire à la traduction de l’uORF aussi situé après. Le mécanisme de régulation pourrait être l’inverse de celui de CAT-1. La traduction de l’uORF, augmentée en l’absence du rG4, nuirait à la structure IRES ou à la traduction IRES-dépendante. Ces hypothèses pourraient être vérifiées en mesurant directement la traduction de l’uORF en présence et en absence du rG4 et en effectuant des essais de gènes rapporteurs bicistroniques en mutant l’uORF pour mesurer son effet sur la traduction IRES-dépendante.

Une étude récente de *ribosome-profiling* avec des cellules HeLa a permis d’identifier les transcrits possédant un rG4 en 5’UTR dont la traduction était réprimée. Ils ont observé que ces transcrits étaient également enrichis pour la présence de uORF en 5’UTR. Le mécanisme proposé est que la présence du rG4 favorise l’initiation à l’uORF réduisant ainsi la traduction du cadre de lecture principal. Ils ont démontré que ce mécanisme était régulé par la présence d’hélicases spécifiques aux rG4, DHX36 et DHX9, qui en dépliant la structure permettent de restaurer la traduction. Ce mécanisme concorde avec les résultats obtenus pour le 5’UTR de BAG-1. Cela est renforcé par l’identification par *pull-down* d’ARN de l’hélicase DHX36 comme partenaire spécifique du rG4 de BAG-1, un résultat obtenu par François Bolduc au laboratoire. En somme, l’effet répresseur des rG4 sur la traduction semble être plus que le résultat d’un encombrement stérique. La corrélation entre la présence de rG4 et de l’uORF suggère un mécanisme plus élaboré où les rG4 favorisent la répression traductionnelle provenant d’uORF.

**Présence de rG4 avec un IRES**

L’impact de la présence d’un rG4 sur la régulation de la traduction dépendante d’un IRES a déjà été étudié pour les ARNm de FGF-2 et de VEGFA (Bonnal *et al.*, 2003 ; Morris *et al.*, 2010 ; Cammas *et al.*, 2015). Le mécanisme d’action du rG4 de BAG-1 est différent puisque contrairement aux exemples précédents le rG4 de BAG-1 ne fait pas partie de la structure secondaire de l’IRES lui-même. Certaines études démontrent aussi l’importance de séquences G-riches qui ont le potentiel d’adopter un rG4 dans des transcrits traduits de façon coiffe indépendante, notamment les 5’UTR des ARNm de l’α-synucléine (SNCA) et de

NRF2 (Koukouraki et Doxakis, 2016 ; Lee *et al.*, 2017). Par contre, dans le cas de SNCA la structure secondaire adoptée n'est pas rigoureusement mesurée, donc la présence d'un rG4 impliqué dans le mécanisme est suggérée, mais non confirmée. Pour NRF2, le rG4 est confirmé, mais la véritable présence d'un IRES n'est pas rigoureusement analysée avec tous les contrôles nécessaires.

Il y a beaucoup de scepticisme concernant l'existence réelle d'IRES chez les eucaryotes. Ces doutes proviennent des nombreuses limites et sources d'erreurs possibles découlant de l'utilisation de gènes rapporteurs bicistroniques et à la fréquente absence de contrôles adéquats (Gilbert, 2010 ; Thompson, 2012 ; Terenin *et al.*, 2017). En premier lieu, il est essentiel de confirmer l'intégrité de la construction bicistronique, c'est-à-dire qu'il faut éliminer les possibilités de promoteurs ou de sites d'épissage cryptiques donnant des transcrits monocistroniques à la suite à la transfection. Dans l'Article 5, ce biais a été écarté par l'analyse d'hybridation Northern pour des sondes spécifiques à la Rluc et la Fluc démontrant une seule longueur de transcrit. De plus, il faut considérer aussi la possibilité de réinitiation à la suite de la traduction du premier cistron qui entraîne la traduction du 2$^{e}$ cistron plutôt qu'une réelle initiation interne due à un IRES.

Dans le cas de BAG-1, la présence d'un IRES ne fait pas de doute, car il été très bien défini et caractérisé dans la littérature (Coldwell *et al.*, 2001 ; Pickering *et al.*, 2004 ; Dobbyn *et al.*, 2008). Dans l'étude présentée ici à l'Article 5, avec l'utilisation d'un plasmide ADN bicistronique avec le 5'UTR complet de BAG-1, la mutation du rG4 entraîne une faible, mais constante, réduction de l'expression du second cistron de 20%. Les essais présentés avec ce type de rapporteur ne permettent pas de discriminer si l'effet est dû à une réduction de la réinitiation ou à une réduction de l'initiation interne à l'IRES. Cependant, puisque la réinitiation et le *leaky scanning* font aussi partie des mécanismes de régulation de la traduction connus de ce transcrit, l'impact de l'absence du rG4 reste pertinent malgré l'incertitude sur le mécanisme à l'origine de cette réduction. C'est plutôt l'analyse de la structure secondaire du 5'UTR complet par SHAPE qui permet de pointer vers une variation de la structure secondaire de l'IRES qui pourrait expliquer la réduction observée.

La lignée cellulaire utilisée, HCT116, est un excellent modèle pour l'étude du cancer colorectal, mais n'est peut-être pas la lignée où l'activité IRES peut être dominante comparativement à d'autres lignées, bien que l'activité IRES dans les HCT116 soit

démontrée dans la littérature à l'aide de rapporteurs bicistroniques également (Lai *et al.*, 2016 ; Wiegering *et al.*, 2015). De plus, les essais ont tous été effectués en conditions de croissance normale. Afin de mieux comprendre les détails de l'impact du rG4, la reproduction de ces essais avec d'autres lignées cellulaires et en conditions de stress connues pour affecter la traduction coiffe-indépendante pourrait nous informer davantage.

**Effet du rG4 sur la structure secondaire globale**

Le mécanisme d'action proposé pour le rG4 de BAG-1 est donc son effet sur la structure secondaire du 5'UTR complet qui a été évaluée grâce à la méthode SHAPE (Article 5). La méthode SHAPE d'évaluation de structure secondaire d'ARN nécessite une étape d'extension d'amorce avec une transcriptase inverse. Or, on sait que les rG4 peuvent stopper cette enzyme, un principe qui est utilisé dans les essais RTS pour étudier les rG4 justement, ce qui empêche donc en général ce type de cartographie. Heureusement, ce problème était contournable pour l'étude du 5'UTR de BAG-1. En effet, puisque le rG4 est situé tout à l'extrémité du 5', l'extension de l'amorce du 3' vers le 5' a pu se faire sur la longueur quasi complète du transcrit. On a donc pu évaluer la structure secondaire qui est en aval du rG4 et qui dans ce cas représente l'ensemble du 5'UTR. D'ailleurs, lors de l'analyse des pics d'électrophorèse capillaire, un large arrêt était présent au G immédiatement en 3' de la dernière série de G ce qui confirme indirectement, encore une fois la présence du rG4.

L'effet du rG4 sur la structure secondaire globale a pu être mesuré en comparant avec une séquence pourtant des mutations G-A dans les séries de G. Bien entendu, ces adénines substituées peuvent former de nouvelles paires de bases et affecter la structure secondaire. Il semble cependant qu'à l'analyse de la structure secondaire obtenue, que les paires de bases de ces adénines restent à proximité et ne forment pas d'interaction de longue portée avec des régions éloignées dans l'UTR, entre autres avec l'IRES. Cependant, pour éviter ce possible biais, des mutations des G avec des 7-déaza-G empêchant les rG4, mais n'affectant pas les autres paires de bases possibles auraient été idéales.

Les résultats de l'Article 5 indiquent que le mécanisme d'action du rG4 sur la traduction cap-indépendante proviendrait de son rôle dans le maintien de la structure secondaire globale puisque sa mutation, sans affecter la stabilité globale prédite de l'UTR, affecte le repliement de sous-éléments clés tel que le site d'entrée du ribosome dans l'IRES. Il faut garder en tête que cette détermination structurale a été effectuée en solution *in vitro*

en absence de tout partenaire protéique. La structure intracellulaire adoptée pourrait différer, ainsi que la présence d'hélicase ou de RBP pourrait modifier le repliement. Il existe à ce jour plusieurs composés de SHAPE qui peuvent traverser la barrière cellulaire, tel que le NAI. Il serait intéressant de répéter l'analyse de la structure secondaire du 5'UTR complet de BAG-1 en condition *in cellulo*. Les études récentes visant à déterminer le repliement rG4 *in cellulo* semblent démontrer que justement ce repliement est dynamique, et non pas présent en tout temps (Al-Zeer et Kurreck, 2018). Afin de mieux décortiquer le mécanisme d'action et l'effet des rG4, l'étude des structures secondaires adoptées *in cellulo* selon les différentes conditions physiologiques favorisant par exemple l'utilisation de l'IRES versus la traduction coiffe-dépendante seront primordiales.

La proposition que le mécanisme d'action d'un rG4 sur la traduction s'effectue par le maintien de la structure secondaire, par l'accessibilité à d'autres séquences régulatrices en *cis* qui permettent d'être liées par des facteurs de traduction et des ITAF commence à gagner de l'attention. Pourtant, ce type de mécanisme est communément accepté et proposé pour expliquer l'effet de structure d'ARN tige-boucle ou d'autres structures secondaires canoniques pour la liaison de RBP en général. Les structures secondaires d'ARN peuvent servir de moyen de mettre en évidence ou de séquestrer des séquences de reconnaissances pour les RBP. Par exemple, une étude démontre que le contexte de la structure secondaire locale où se retrouve le motif de reconnaissance est tout aussi important que la séquence elle-même pour être reconnu par la RBP (Taliaferro *et al.*, 2016).

De plus, le 5'UTR de BAG-1 n'est pas l'unique exemple où la formation d'un rG4 en conjonction avec d'autres éléments affecte la traduction. La présence d'un rG4 et d'une tige-boucle dans le 5'UTR de l'ARNm résulte en l'inhibition de la synthèse protéique de l'un des 12 isoformes codés par le gène HNF4A de la souris, précisément l'isoforme alpha1. Individuellement, le rG4 ou la tige-boucle ne répriment pas l'expression (Guo et Lu, 2017, 2018). L'effet d'un rG4 a été démontré comme étant fortement dépendant du contexte en effectuant un « échange » de deux rG4 aux effets opposés (un réprime la traduction et le second active la traduction). Même échangé, c'est l'effet de l'UTR initial qui était observé démontrant que c'est le contexte et probablement la structure secondaire globale adoptée en présence d'un rG4 qui sont importants et non pas le rG4 en lui-même (Bhattacharyya *et al.*, 2017). Il faut donc changer le paradigme où la structure rG4 agit seule comme un « bloc ».

Au contraire, tout indique que c'est une structure secondaire dynamique qui peut s'alterner avec des structures secondaires canoniques et contribuer au repliement global et que c'est cette dynamique qui joue probablement le plus fortement sur la régulation.

Individuellement, chacun des quatre éléments en *cis* abordés, les codons de départ alternatifs, les codons de départ non canoniques, les uORF et les IRES, influencent la traduction des ARNm. L'originalité des travaux de cette thèse provient de la considération simultanée de tous ces éléments dans un contexte adjacent à un rG4 avec l'étude du candidat 5'UTR de BAG1. En conclusion, les éléments du contexte entourant le rG4 et non seulement sa stabilité intrinsèque sont importants dans le mécanisme régulant la traduction.

# CONCLUSION

Les travaux présentés dans cette thèse ont permis d'étendre nos connaissances sur les structures G4 d'ARN. En effet, l'utilisation de la cartographie *in-line* adaptée au rG4 sur des séquences d'ARN très longues a permis d'étudier l'impact du contexte nucléotidique sur le repliement de la structure. Des séquences naturelles issues directement d'ARNm humain, variées en termes de longueurs de contextes ainsi qu'en motifs PG4 ont été cartographiées. Cette méthode *in vitro* qui respecte mieux les conditions physiologiques que les méthodes traditionnelles d'étude rG4 a permis d'examiner de façon plus systématique l'impact des séquences extérieures au motif canonique sur le repliement. Cela a permis d'établir un système de score simple, mais efficace qui permet de mieux prédire la formation rG4 en tenant compte des C et des G consécutifs présents dans le contexte des motifs PG4. Ces deux outils peuvent être appliqués à l'ensemble du transcriptome afin de déterminer les séquences susceptibles d'adopter des rG4. Ils ont permis d'identifier plusieurs nouveaux candidats rG4 présents dans des ARNm reliés à des voies biologiques, dont le cancer colorectal.

La cartographie *in-line* permet aussi de démontrer le repliement de rG4 irréguliers, dont ceux avec de longues boucles centrales. Elle permet aussi d'observer des caractéristiques plus spécifiques des rG4 comme la présence et la composition de plusieurs boucles possibles ainsi que l'ensemble des séries de G impliquées dans les multiples conformations rG4 d'une seule séquence. Cela permet d'identifier des caractéristiques communes aux rG4 comme ceux retrouvés dans les 5'UTR d'ARNm associés au cancer colorectal. Cela suggère que des sous-groupes de rG4 particuliers pourraient être impliqués dans la corégulation de l'expression d'ARNm.

Les rG4 sont impliqués dans de nombreux processus de régulation post-transcriptionnels. Leurs contributions sont particulièrement étudiées dans la traduction des ARNm où les rG4 en 5'UTR agissent en tant que répresseurs. Les travaux de cette thèse avec l'étude du candidat rG4 du 5'UTR de BAG-1 ont permis d'étendre la compréhension du mécanisme d'action du rG4 sur la traduction. Un seul rG4 peut avoir des effets opposés sur la traduction coiffe-dépendante et coiffe-indépendante. Cela a de plus permis de constater l'importance de l'interaction du rG4 avec plusieurs autres éléments régulateurs situés dans son contexte nucléotidique comme des codons de départ alternatifs, un uORF et une structure

IRES. De plus, l'impact du rG4 au niveau de la structure secondaire globale du 5'UTR a été démontré.

Les rG4 sont abondants et ont des effets directs sur la structure secondaire globale et la régulation post-transcriptionnelle. Les méthodes de prédiction bio-informatique du score cG/cC et d'étude *in vitro* par cartographie *in-line* des séquences rG4 variées avec un large contexte nucléotidique qui ont été présentées dans cette thèse pourront être appliquées pour investiguer le rôle de bien de nouveaux rG4 situés dans une grande variété de familles ou de localisations de l'ARN afin de mieux comprendre l'éventail de leurs fonctions biologiques.

# LISTE DES RÉFÉRENCES

Adrian, M., Heddi, B. et Phan, A. T., (2012), NMR spectroscopy of G-quadruplexes, *Methods*, vol. 57, n° 1, p. 11-24.

Agarwala, P., Pandey, S. et Maiti, S., (2014), Role of G-quadruplex located at 5' end of mRNAs., *Biochim. Biophys. Acta*, vol. 1840, p. 3503-3510.

Agarwala, P., Pandey, S., Mapa, K. et Maiti, S., (2013), The G-Quadruplex Augments Translation in the 5′ Untranslated Region of Transforming Growth Factor β2, *Biochemistry*, vol. 52, n°9, p. 1528-1538.

Agrawal, P., Hatzakis, E., Guo, K., Carver, M. et Yang, D., (2013), Solution structure of the major G-quadruplex formed in the human VEGF promoter in K+: insights into loop interactions of the parallel G-quadruplexes, *Nucleic Acids Res.*, vol. 41, n°22, p. 10584-10592.

Alberti, S., Esser, C. et Höhfeld, J., (2003), BAG-1--a nucleotide exchange factor of Hsc70 with multiple cellular functions, *Cell Stress Chaperones*, vol. 8, n°3, p. 225-231.

Al-Zeer, M. A. et Kurreck, J., (2018), Deciphering the Enigmatic Biological Functions of RNA Guanine-Quadruplex Motifs in Human Cells, *Biochemistry*.

Amor, S., Yang, S. Y., Wong, J. M. Y. et Monchaud, D., (2017), Cellular Detection of G-Quadruplexes by Optical Imaging Methods, *Curr. Protoc. Cell Biol.*, vol. 76, p. 4.33.1-4.33.19.

Amrane, S., Adrian, M., Heddi, B., Serero, A., Nicolas, A., Mergny, J.-L. et Phan, A. T., (2012), Formation of pearl-necklace monomorphic G-quadruplexes in the human CEB25 minisatellite, *J. Am. Chem. Soc.*, vol. 134, n°13, p. 5807-5816.

Arcondéguy, T., Lacazette, E., Millevoi, S., Prats, H. et Touriol, C., (2013), VEGF-A mRNA processing, stability and translation: a paradigm for intricate regulation of gene expression at the post-transcriptional level, *Nucleic Acids Res.*, vol. 41, n°17, p. 7997-8010.

Ariyo, E. O., Booy, E. P., Dzananovic, E., McRae, E. K., Meier, M., McEleney, K., … McKenna, S. A., (2017), Impact of G-quadruplex loop conformation in the PITX1 mRNA on protein and small molecule interaction, *Biochem. Biophys. Res. Commun.*, vol. 487, n°2, p. 274-280.

Armas, P. et Calcaterra, N. B., (2018), G-quadruplex in animal development : Contribution to gene expression and genomic heterogeneity, *Mech. Dev.*, vol. 154, p. 4-72.

Arora, A., Dutkiewicz, M., Scaria, V., Hariharan, M., Maiti, S. et Kurreck, J., (2008), Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif., *RNA*, vol. 14, p. 1290-1296.

Arora, A., Nair, D. R. et Maiti, S., (2009), Effect of flanking bases on quadruplex stability and Watson-Crick duplex competition, *FEBS J.*, vol. 276, n°13, p. 3628-3640.

Arora, A. et Suess, B., (2011), An RNA G-quadruplex in the 3' UTR of the proto-oncogene PIM1 represses translation, *RNA Biol.*, vol. 8, p. 802-805.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G., (2000), Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.*, vol. 25, n°1, p. 25-29.

Aveic, S., Viola, G., Accordi, B., Micalizzi, C., Santoro, N., Masetti, R., … Pigazzi, M., (2015), Targeting BAG-1: a novel strategy to increase drug efficacy in acute myeloid leukemia, *Exp. Hematol.*, vol. 43, n°3, p. 180-190.e6.

Babendure, J. R., Babendure, J. L., Ding, J.-H. et Tsien, R. Y., (2006), Control of mammalian translation by mRNA structure near caps, *RNA*, vol. 12, n°5, p. 851-861.

Balkwill, D. G., Derecka, K., Garner, P. T., Hodgman, C., Flint, F. A. P. et Searle, S. M., (2009), Repression of translation of human estrogen receptor alpha by G-quadruplex formation., *Biochemistry*, vol. 48, p. 11487-11495.

Bao, H.-L. et Xu, Y., (2018), Investigation of higher-order RNA G-quadruplex structures *in vitro* and in living cells by $^{19}$F NMR spectroscopy, *Nat. Protoc.*, vol. 13, n°4, p. 652-665.

Baral, A., Kumar, P., Halder, R., Mani, P., Yadav, V. K., Singh, A., … Chowdhury, S., (2012), Quadruplex-single nucleotide polymorphisms (Quad-SNP) influence gene expression difference among individuals, *Nucleic Acids Res.*, vol. 40, n°9, p. 3800-3811.

Barbosa, C., Peixeiro, I. et Romão, L., (2013), Gene Expression Regulation by Upstream Open Reading Frames and Human Disease, *PLoS Genet.*, vol. 9, n°8, p. e1003529.

Barnes, J. D., Arhel, N. J., Lee, S. S., Sharp, A., Al-Okail, M., Packham, G., … Williams, A. C., (2005), Nuclear BAG-1 expression inhibits apoptosis in colorectal adenoma-derived epithelial cells, *Apoptosis*, vol. 10, n°2, p. 301-311.

Bastide, A., Karaa, Z., Bornes, S., Hieblot, C., Lacazette, E., Prats, H. et Touriol, C., (2008), An upstream open reading frame within an IRES controls expression of a specific VEGF-A isoform, *Nucleic Acids Res.*, vol. 36, n°7, p. 2434-2445.

Beaudoin, J.-D., Jodoin, R. et Perreault, J.-P., (2013), In-line probing of RNA G-quadruplexes., *Methods*, vol. 64, p. 79-87.

Beaudoin, J.-D., Jodoin, R. et Perreault, J.-P., (2014), New scoring system to identify RNA G-quadruplex folding., *Nucleic Acids Res.*, vol. 42, n°2, p. 1209-1223.

Beaudoin, J.-D. et Perreault, J.-P., (2010), 5'-UTR G-quadruplex structures acting as translational repressors., *Nucleic Acids Res.*, vol. 38, n°20, p. 7022-7036.

Beaudoin, J.-D. et Perreault, J.-P., (2013), Exploring mRNA 3'-UTR G-quadruplexes : evidence of roles in both alternative polyadenylation and mRNA shortening., *Nucleic Acids Res.*, vol. 41, n°11, p. 5898-5911.

Bedrat, A., Lacroix, L. et Mergny, J.-L., (2016), Re-evaluation of G-quadruplex propensity with G4Hunter, *Nucleic Acids Res.*, vol. 44, n°4, p. 1746-1759.

Benhalevy, D., Gupta, S. K., Danan, C. H., Ghosal, S., Sun, H.-W., Kazemier, H. G., … Juranek, S. A., (2017), The Human CCHC-type Zinc Finger Nucleic Acid-Binding Protein Binds G-Rich Elements in Target mRNA Coding Sequences and Promotes Translation, *Cell Rep.*, vol. 18, n°12, p. 2979-2990.

Bensaid, M., Melko, M., Bechara, E. G., Davidovic, L., Berretta, A., Catania, M. V., … Bardoni, B., (2009), FRAXE-associated mental retardation protein (FMR2) is an RNA-binding protein with high affinity for G-quartet RNA forming structure, *Nucleic Acids Res.*, vol. 37, n°4, p. 1269-1279.

Bhasikuttan, A. C. et Mohanty, J., (2015), Targeting G-quadruplex structures with extrinsic fluorogenic dyes : promising fluorescence sensors, *Chem. Commun.*, vol. 51, n°36, p. 7581-7597.

Bhattacharyya, D., Diamond, P. et Basu, S., (2015), An Independently folding RNA G-quadruplex domain directly recruits the 40S ribosomal subunit., *Biochemistry*, vol. 54, p. 1879-1885.

Bhattacharyya, D., Morris, M. J., Kharel, P., Mirihana Arachchilage, G., Fedeli, K. M. et Basu, S., (2017), Engineered domain swapping indicates context dependent functional role of RNA G-quadruplexes, *Biochimie*, vol. 137, p. 147-150.

Bian, B., Mongrain, S., Cagnol, S., Langlois, M., Boulanger, J., Bernatchez, G., … Rivard, N., (2016), Cathepsin B promotes colorectal tumorigenesis, cell invasion, and metastasis, *Mol. Carcinog.*, vol. 55, n°5, p. 671-687.

Bian, Y., Tan, C., Wang, J., Sheng, Y., Zhang, J. et Wang, W., (2014), Atomistic picture for the folding pathway of a hybrid-1 type human telomeric DNA G-quadruplex, *PLoS Comput. Biol.*, vol. 10, n°4, p. e1003562.

Biffi, G., Di Antonio, M., Tannahill, D. et Balasubramanian, S., (2014a), Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells, *Nat. Chem.*, vol. 6, n°1, p. 75-80.

Biffi, G., Tannahill, D. et Balasubramanian, S., (2012), An intramolecular G-quadruplex structure is required for binding of telomeric repeat-containing RNA to the telomeric protein TRF2., *J. Am. Chem. Soc.*, vol. 134, n°29, p. 11974-11976.

Biffi, G., Tannahill, D., McCafferty, J. et Balasubramanian, S., (2013), Quantitative visualization of DNA G-quadruplex structures in human cells, *Nat. Chem.*, vol. 5, n°3, p. 182-186.

Biffi, G., Tannahill, D., Miller, J., Howat, W. J. et Balasubramanian, S., (2014b), Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues, *PLoS One*, vol. 9, n°7, p. e102711.

Bolduc, F., Garant, J.-M., Allard, F. et Perreault, J.-P., (2016), Irregular G-quadruplexes Found in the Untranslated Regions of Human mRNAs Influence Translation, *J. Biol. Chem.*, vol. 291, n°41, p. 21751-21760.

Bonnal, S., Schaeffer, C., Créancier, L., Clamens, S., Moine, H., Prats, A.-C. et Vagner, S., (2003), A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons, *J. Biol. Chem.*, vol. 278, n°41, p. 39330-39336.

Booy, P. E., McRae, S. E. K., Howard, R., Deo, R. S., Ariyo, O. E., Dzananovic, E., … McKenna, A. S., (2016), RNA Helicase Associated with AU-rich Element (RHAU/DHX36) Interacts with the 3'-Tail of the Long Non-coding RNA BC200 (BCYRN1)., *J. Biol. Chem.*, vol. 291, n°10, p. 5355-5372.

Booy, P. E., Meier, M., Okun, N., Novakowski, K. S., Xiong, S., Stetefeld, J. et McKenna, A. S., (2012), The RNA helicase RHAU (DHX36) unwinds a G4-quadruplex in human telomerase RNA and promotes the formation of the P1 helix template boundary., *Nucleic Acids Res.*, vol. 40, n°9, p. 4110-4124.

Bouchard, P. et Legault, P., (2014), Structural insights into substrate recognition by the Neurospora Varkud satellite ribozyme: importance of U-turns at the kissing-loop junction, *Biochemistry*, vol. 53, n°1, p. 258-269.

Brázda, V., Červeň, J., Bartas, M., Mikysková, N., Coufal, J., Pečinka, P., … Pečinka, P., (2018), The Amino Acid Composition of Quadruplex Binding Proteins Reveals a Shared Motif and Predicts New Potential Quadruplex Interactors, *Molecules*, vol. 23, n°9, p. 2341.

Brosseau, J.-P., Lucier, J.-F., Lapointe, E., Durand, M., Gendron, D., Gervais-Bird, J., … Elela, S. A., (2010), High-throughput quantification of splicing isoforms, *RNA*, vol. 16, n°2, p. 442-449.

Bugaut, A. et Balasubramanian, S., (2008), A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes, *Biochemistry*, vol. 47, n°2, p. 689-697.

Bugaut, A. et Balasubramanian, S., (2012), 5'-UTR RNA G-quadruplexes : translation regulation and targeting., *Nucleic Acids Res.*, vol. 40, n°11, p. 4727-4741.

Bugaut, A., Murat, P. et Balasubramanian, S., (2012), An RNA hairpin to G-quadruplex conformational transition., *J. Am. Chem. Soc.*, vol. 134, n°49, p. 19953-19956.

Bugaut, A., Rodriguez, R., Kumari, S., Hsu, S.-T. D. et Balasubramanian, S., (2010), Small molecule-mediated inhibition of translation by targeting a native RNA G-quadruplex, *Org. Biomol. Chem.*, vol. 8, n°12, p. 2771-2776.

Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K. et Neidle, S., (2006), Quadruplex DNA : sequence, topology and structure, *Nucleic Acids Res.*, vol. 34, n°19, p. 5402-5415.

Calkhoven, C. F., Müller, C. et Leutz, A., (2000), Translational control of C/EBPα and C/EBPβ isoform expression, *Genes Dev.*, vol. 14, n°15, p. 1920-1932.

Calvo, S. E., Pagliarini, D. J. et Mootha, V. K., (2009), Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, n°18, p. 7507-7512.

Cammas, A., Dubrac, A., Morel, B., Lamaa, A., Touriol, C., Teulade-Fichou, M.-P., … Millevoi, S., (2015), Stabilization of the G-quadruplex at the VEGF IRES represses cap-independent translation, *RNA Biol.*, vol. 12, n°3, p. 320-329.

Cammas, A., Lacroix-Triki, M., Pierredon, S., Le Bras, M., Iacovoni, S., Jason, Teulade-Fichou, M.-P., … Vagner, S., (2016), hnRNP A1-mediated translational regulation of the G quadruplexcontaining RON receptor tyrosine kinase mRNA linked to tumor progression, *Oncotarget*, vol. 7, n°13, p. 16793–16805.

Cammas, A. et Millevoi, S., (2017), RNA G-quadruplexes: emerging mechanisms in disease, *Nucleic Acids Res.*, vol. 45, n°4, p. 1584-1595.

Cancer Genome Atlas Network, (2012), Comprehensive molecular characterization of human colon and rectal cancer, *Nature*, vol. 487, n°7407, p. 330-337.

Capra, J. A., Paeschke, K., Singh, M. et Zakian, V. A., (2010), G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae, *PLoS Comput. Biol.*, vol. 6, n°7, p. e1000861.

Cato, L., Neeb, A., Sharp, A., Buzón, V., Ficarro, S. B., Yang, L., … Brown, M., (2017), Development of Bag-1L as a therapeutic target in androgen receptor-dependent prostate cancer, *ELife*, vol. 6, p. e27159.

Cavaliere, P., Pagano, B., Granata, V., Prigent, S., Rezaei, H., Giancola, C. et Zagari, A., (2013), Cross-talk between prion protein and quadruplex-forming nucleic acids: a dynamic complex formation, *Nucleic Acids Res.*, vol. 41, n°1, p. 327-339.

Cencic, R., Galicia-Vázquez, G. et Pelletier, J., (2012), Chapter Twenty - Inhibitors of Translation Targeting Eukaryotic Translation Initiation Factor 4A, In : E. JANKOWSKY (éd.), *Methods in Enzymology*, Academic Pressvol. 511, , p. 437-461.

Cer, R. Z., Bruce, K. H., Donohue, D. E., Temiz, N. A., Mudunuri, U. S., Yi, M., … Stephens, R. M., (2012), Searching for non-B DNA-forming motifs using nBMST (non-B DNA motif search tool), *Curr. Protoc. Hum. Genet.*, vol. Chapter 18, p. Unit 18.7.1-22.

Chakraborty, P. et Grosse, F., (2011), Human DHX9 helicase preferentially unwinds RNA-containing displacement loops (R-loops) and G-quadruplexes, *DNA Repair*, vol. 10, n°6, p. 654-665.

Chatterji, P. et Rustgi, A. K., (2018), RNA Binding Proteins in Intestinal Epithelial Biology and Colorectal Cancer, *Trends Mol Med*, vol. 24, n°5, p. 490-506.

Chen, M. C., Tippana, R., Demeshkina, N. A., Murat, P., Balasubramanian, S., Myong, S. et Ferré-D'Amaré, A. R., (2018a), Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36, *Nature*, vol. 558, n°7710, p. 465-469.

Chen, X.-C., Chen, S.-B., Dai, J., Yuan, J.-H., Ou, T.-M., Huang, Z.-S. et Tan, J.-H., (2018b), Tracking the Dynamic Folding and Unfolding of RNA G-Quadruplexes in Live Cells, *Angew. Chem. Int. Ed. Engl.*, vol. 57, n°17, p. 4702-4706.

Cheng, M., Cheng, Y., Hao, J., Jia, G., Zhou, J., Mergny, J.-L. et Li, C., (2018), Loop permutation affects the topology and stability of G-quadruplexes, *Nucleic Acids Res.*, vol. 46, n°18, p. 9264-9275.

Cho, H., Cho, H. S., Nam, H., Jo, H., Yoon, J., Park, C., … Hwang, I., (2018), Translational control of phloem development by RNA G-quadruplex–JULGI determines plant sink strength, *Nat. Plants*, vol. 4, n°6, p. 376-390.

Christiansen, J., Kofod, M. et Nielsen, C. F., (1994), A guanosine quadruplex and two stable hairpins flank a major cleavage site in insulin-like growth factor II mRNA., *Nucleic Acids Res.*, vol. 22, n°25, p. 5709-5716.

Cian, A. D., Cristofari, G., Reichenbach, P., Lemos, E. D., Monchaud, D., Teulade-Fichou, M.-P., … Mergny, J.-L., (2007), Reevaluation of telomerase inhibition by quadruplex ligands and their mechanisms of action, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, n°44, p. 17347-17352.

Clemo, N. K., Arhel, N. J., Barnes, J. D., Baker, J., Moorghen, M., Packham, G. K., … Williams, A. C., (2005), The role of the retinoblastoma protein (Rb) in the nuclear localization of BAG-1: implications for colorectal tumour cell survival, *Biochem. Soc. Trans.*, vol. 33, n°Pt 4, p. 676-678.

Clemo, N. K., Collard, T. J., Southern, S. L., Edwards, K. D., Moorghen, M., Packham, G., … Williams, A. C., (2008), BAG-1 is up-regulated in colorectal tumour progression and promotes colorectal tumour cell survival through increased NF-kappaB activity, *Carcinogenesis*, vol. 29, n°4, p. 849-857.

Cogoi, S. et Xodo, L. E., (2006), G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription, *Nucleic Acids Res.*, vol. 34, n°9, p. 2536-2549.

Coldwell, M. J., deSchoolmeester, M. L., Fraser, G. A., Pickering, B. M., Packham, G. et Willis, A. E., (2001), The p36 isoform of BAG-1 is translated by internal ribosome entry following heat shock, *Oncogene*, vol. 20, n°30, p. 4095-4100.

Collard, T. J., Urban, B. C., Patsos, H. A., Hague, A., Townsend, P. A., Paraskeva, C. et Williams, A. C., (2012), The retinoblastoma protein (Rb) as an anti-apoptotic factor:

expression of Rb is required for the anti-apoptotic function of BAG-1 protein in colorectal tumour cells, *Cell Death Dis.*, vol. 3, p. e408.

Collie, G. W., Haider, S. M., Neidle, S. et Parkinson, G. N., (2010), A crystallographic and modelling study of a human telomeric RNA (TERRA) quadruplex, *Nucleic Acids Res.*, vol. 38, n°16, p. 5569-5580.

Collie, G. W. et Parkinson, G. N., (2011), The application of DNA and RNA G-quadruplexes to therapeutic medicines, *Chem. Soc. Rev.*, vol. 40, n°12, p. 5867-5892.

Conlon, E. G., Lu, L., Sharma, A., Yamazaki, T., Tang, T., Shneider, N. A. et Manley, J. L., (2016), The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains, *ELife*, vol. 5, p. e17820.

Corley, M., Solem, A., Phillips, G., Lackey, L., Ziehr, B., Vincent, H. A., … Laederach, A., (2017), An RNA structure-mediated, posttranscriptional model of human α-1-antitrypsin expression, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, n°47, p. E10244-E10253.

Costantino, L. et Koshland, D., (2015), The Yin and Yang of R-loop Biology, *Curr. Opin. Cell Biol.*, vol. 34, p. 39-45.

Crappé, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., … Menschaert, G., (2015), PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration, *Nucleic Acids Res.*, vol. 43, n°5, p. e29.

Creacy, S. D., Routh, E. D., Iwamoto, F., Nagamine, Y., Akman, S. A. et Vaughn, J. P., (2008), G4 Resolvase 1 Binds Both DNA and RNA Tetramolecular Quadruplex with High Affinity and Is the Major Source of Tetramolecular Quadruplex G4-DNA and G4-RNA Resolving Activity in HeLa Cell Lysates, *J. Biol. Chem.*, vol. 283, n°50, p. 34626-34634.

Crenshaw, E., Leung, P. B., Kwok, K. C., Sharoni, M., Olson, K., Sebastian, P. N., … Saunders, J. A., (2015), Amyloid Precursor Protein Translation Is Regulated by a 3'UTR Guanine Quadruplex., *PloS One*, vol. 10, n°11, p. e0143160.

Crick, F., (1970), Central Dogma of Molecular Biology, *Nature*, vol. 227, n°5258, p. 561.

Cruz, J. A. et Westhof, E., (2009), The Dynamic Landscapes of RNA Architecture, *Cell*, vol. 136, n°4, p. 604-609.

Darty, K., Denise, A. et Ponty, Y., (2009), VARNA : Interactive drawing and editing of the RNA secondary structure, *Bioinformatics*, vol. 25, n°15, p. 1974-1975.

Das, K., Srivastava, M. et Raghavan, S. C., (2016), GNG Motifs Can Replace a GGG Stretch during G-Quadruplex Formation in a Context Dependent Manner, *PLoS One*, vol. 11, n°7, p. e0158794.

Das, R., Laederach, A., Pearlman, S. M., Herschlag, D. et Altman, R. B., (2005), SAFA : semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments, *RNA*, vol. 11, n°3, p. 344-354.

De Cian, A., DeLemos, E., Mergny, J.-L., Teulade-Fichou, M.-P. et Monchaud, D., (2007), Highly Efficient G-Quadruplex Recognition by Bisquinolinium Compounds, *J. Am. Chem. Soc.*, vol. 129, n°7, p. 1856-1857.

Decorsiere, A., Cayrel, A., Vagner, S. et Millevoi, S., (2011), Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3'-end processing and function during DNA damage., *Genes Dev.*, vol. 25, n°3, p. 220-225.

Del Villar-Guerra, R., Trent, J. O. et Chaires, J. B., (2018), G-quadruplex secondary structure from circular dichroism spectroscopy, *Angew. Chem., Int. Ed. Engl.*, vol. 57, n°24, p. 7171-7175.

Dempsey, L. A., Sun, H., Hanakahi, L. A. et Maizels, N., (1999), G4 DNA binding by LR1 and its subunits, nucleolin and hnRNP D, A role for G-G pairing in immunoglobulin switch recombination, *J. Biol. Chem.*, vol. 274, n°2, p. 1066-1071.

Dhapola, P. et Chowdhury, S., (2016), QuadBase2: web server for multiplexed guanine quadruplex mining and visualization, *Nucleic Acids Res.*, vol. 44, n°Web Server issue, p. W277-W283.

Di Antonio, M., Biffi, G., Mariani, A., Raiber, E.-A., Rodriguez, R. et Balasubramanian, S., (2012), Selective RNA versus DNA G-quadruplex targeting by in situ click chemistry, *Angew. Chem., Int. Ed. Engl.*, vol. 51, n°44, p. 11073-11078.

Di Leva, F. S., Novellino, E., Cavalli, A., Parrinello, M. et Limongelli, V., (2014), Mechanistic insight into ligand binding to G-quadruplex DNA, *Nucleic Acids Res.*, vol. 42, n°9, p. 5447-5455.

Diaz de Arce, A. J., Noderer, W. L. et Wang, C. L., (2018), Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons, *Nucleic Acids Res.*, vol. 46, n°2, p. 985-994.

Didiot, M.-C., Tian, Z., Schaeffer, C., Subramanian, M., Mandel, J.-L. et Moine, H., (2008), The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer., *Nucleic Acids Res.*, vol. 36, n°15, p. 4902-4912.

Ding, Y., Fleming, A. M. et Burrows, C. J., (2018), Case studies on potential G-quadruplex-forming sequences from the bacterial orders Deinococcales and Thermales derived from a survey of published genomes, *Sci. Rep.*, vol. 8, n°1, p. 15679.

Dobbyn, H. C., Hill, K., Hamilton, T. L., Spriggs, K. A., Pickering, B. M., Coldwell, M. J., … Willis, A. E., (2008), Regulation of BAG-1 IRES-mediated translation following chemotoxic stress, *Oncogene*, vol. 27, n°8, p. 1167-1174.

Du, Z., Yu, J., Ulyanov, N. B., Andino, R. et James, T. L., (2004), Solution structure of a consensus stem-loop D RNA domain that plays important roles in regulating translation and replication in enteroviruses and rhinoviruses, *Biochemistry*, vol. 43, n°38, p. 11959-11972.

Eddy, J. et Maizels, N., (2006), Gene function correlates with potential for G4 DNA formation in the human genome, *Nucleic Acids Res.*, vol. 34, n°14, p. 3887-3896.

Eddy, J. et Maizels, N., (2008), Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes, *Nucleic Acids Res*, vol. 36, n°4, p. 1321-1333.

ENCODE Project Consortium, (2007), Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, vol. 447, n°7146, p. 799-816.

Endoh, T., Kawasaki, Y. et Sugimoto, N., (2013a), Stability of RNA quadruplex in open reading frame determines proteolysis of human estrogen receptor α., *Nucleic Acids Res.*, vol. 41, n°12, p. 6222-6231.

Endoh, T., Kawasaki, Y. et Sugimoto, N., (2013b), Suppression of gene expression by G-quadruplexes in open reading frames depends on G-quadruplex stability., *Angew. Chem., Int. Ed. Engl.*, vol. 52, n°21, p. 5522-5526.

Endoh, T., Kawasaki, Y. et Sugimoto, N., (2013c), Translational halt during elongation caused by G-quadruplex formed by mRNA., *Methods*, vol. 64, n°1, p. 73-78.

Endoh, T. et Sugimoto, N., (2013), Unusual -1 ribosomal frameshift caused by stable RNA G-quadruplex in open reading frame., *Anal. Chem.*, vol. 85, n°23, p. 11435-11439.

Endoh, T. et Sugimoto, N., (2016), Mechanical insights into ribosomal progression overcoming RNA G-quadruplex from periodical translation suppression in cells., *Sci. Rep.*, vol. 6, p. 22719.

Fay, M. M., Lyons, S. M. et Ivanov, P., (2017), RNA G-Quadruplexes in Biology : Principles and Molecular Mechanisms, *J. Mol. Biol.*, vol. 429, n°14, p. 2127-2147.

Fisette, J.-F., Montagna, R. D., Mihailescu, M.-R. et Wolfe, S. M., (2012), A G-rich element forms a G-quadruplex and regulates BACE1 mRNA alternative splicing., *J. Neurochem.*, vol. 121, n°5, p. 763-773.

Fleming, A. M., Zhu, J., Ding, Y., Visser, J. A., Zhu, J. et Burrows, C. J., (2018), Human DNA Repair Genes Possess Potential G-Quadruplex Sequences in Their Promoters and 5′-Untranslated Regions, *Biochemistry*, vol. 57, n°6, p. 991-1002.

Fratta, P., Mizielinska, S., Nicoll, A. J., Zloh, M., Fisher, E. M. C., Parkinson, G. et Isaacs, A. M., (2012), *C9orf72* hexanucleotide repeat associated with amyotrophic lateral sclerosis and frontotemporal dementia forms RNA G-quadruplexes, *Sci. Rep.*, vol. 2, p. 1016.

Frees, S., Menendez, C., Crum, M. et Bagga, P. S., (2014), QGRS-Conserve : a computational method for discovering evolutionarily conserved G-quadruplex motifs, *Hum. Genomics*, vol. 8, p. 8.

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., … Brosch, M., (2012), Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting, *Genome Res.*, vol. 22, n°11, p. 2208-2218.

Fuller-Pace, F. V., (2013), DEAD box RNA helicase functions in cancer, *RNA Biol.*, vol. 10, n°1, p. 121-132.

Galloway, A. et Cowling, V. H., (2018), mRNA cap regulation in mammalian cell function and fate, *Biochim. Biophys. Acta, Gene Regul. Mech.*, vol. 18, p. 30167-6.

Garant, J.-M., Luce, M. J., Scott, M. S. et Perreault, J.-P., (2015), G4RNA : an RNA G-quadruplex database, *Database (Oxford)*, vol. 2015, p. bav059.

Garant, J.-M., Perreault, J.-P. et Scott, M. S., (2017), Motif independent identification of potential RNA G-quadruplexes by G4RNA screener, *Bioinformatics*, vol. 33, n°22, p. 3532-3537.

Garant, J.-M., Perreault, J.-P. et Scott, M. S., (2018), G4RNA screener web server : User focused interface for RNA G-quadruplex prediction, *Biochimie*, vol. 151, p. 115-118.

Garg, R., Aggarwal, J. et Thakkar, B., (2016), Genome-wide discovery of G-quadruplex forming sequences and their functional relevance in plants, *Sci. Rep.*, vol. 6, p. 28211.

Gellert, M., Lipsett, M. N. et Davies, D. R., (1962), Helix formation by guanylic acid, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 48, p. 2013-2018.

Ghosh, A., Ekka, M. K., Tawani, A., Kumar, A., Chakraborty, D. et Maiti, S., (2018), Restoration of miRNA-149 expression by TmPyP4 induced unfolding of quadruplex within its precursor, *Biochemistry*.

Giguère, T. et Perreault, J.-P., (2017), Classification of the Pospiviroidae based on their structural hallmarks, *PloS One*, vol. 12, n°8, p. e0182536.

Gilbert, W. V., (2010), Alternative Ways to Think about Cellular Internal Ribosome Entry, *J. Biol. Chem.*, vol. 285, n°38, p. 29033-29038.

Glouzon, J.-P. S., Perreault, J.-P. et Wang, S., (2017a), Structurexplor: a platform for the exploration of structural features of RNA secondary structures, *Bioinformatics*, vol. 33, n°19, p. 3117-3120.

Glouzon, J.-P. S., Perreault, J.-P. et Wang, S., (2017b), The super-n-motifs model : a novel alignment-free approach for representing and comparing RNA secondary structures, *Bioinformatics*, vol. 33, n°8, p. 1169-1178.

Gomez, D., Guedin, A., Mergny, J.-L., Salles, B., Riou, J.-F., Teulade-Fichou, M.-P. et Calsou, P., (2010), A G-quadruplex structure within the 5'-UTR of TRF2 mRNA represses translation in human cells, *Nucleic Acids Res.*, vol. 38, n°20, p. 7187-7198.

Gomez, D., Lemarteleur, T., Lacroix, L., Mailliet, P., Mergny, J.-L. et Riou, J.-F., (2004), Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing., *Nucleic Acids Res.*, vol. 32, n°1, p. 371-379.

Gray, R. D., Pettracone, L., Buscaglia, R. et Chaires, J. B., (2010), 2-Aminopurine as a Probe for Quadruplex Loop Structures, *Methods Mol Biol*, vol. 608, p. 121-136.

Gros, J., Guédin, A., Mergny, J.-L. et Lacroix, L., (2008), G-Quadruplex formation interferes with P1 helix formation in the RNA component of telomerase hTERC., *Chembiochem*, vol. 9, n°13, p. 2075-2079.

Gros, J., Rosu, F., Amrane, S., De Cian, A., Gabelica, V., Lacroix, L. et Mergny, J.-L., (2007), Guanines are a quartet's best friend : impact of base substitutions on the kinetics and stability of tetramolecular quadruplexes, *Nucleic Acids Res.*, vol. 35, n°9, p. 3064-3075.

Guédin, A., Alberti, P. et Mergny, J.-L., (2009), Stability of intramolecular quadruplexes : sequence effects in the central loop, *Nucleic Acids Res.*, vol. 37, n°16, p. 5559-5567.

Guédin, A., De Cian, A., Gros, J., Lacroix, L. et Mergny, J.-L., (2008), Sequence effects in single-base loops for quadruplexes, *Biochimie*, vol. 90, n°5, p. 686-696.

Guédin, A., Gros, J., Alberti, P. et Mergny, J.-L., (2010), How long is too long? Effects of loop size on G-quadruplex stability, *Nucleic Acids Res.*, vol. 38, n°21, p. 7858-7868.

Guo, J. U. et Bartel, D. P., (2016), RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria, *Science*, vol. 353, n°6306, p. aaf5371.

Guo, S. et Lu, H., (2017), Conjunction of potential G-quadruplex and adjacent cis-elements in the 5′ UTR of hepatocyte nuclear factor 4-alpha strongly inhibit protein expression, *Sci. Rep.*, vol. 7, n°1, p. 17444.

Guo, S. et Lu, H., (2018), Conjunction of G-quadruplex and stem-loop in the 5' untranslated region of mouse hepatocyte nuclear factor 4-alpha1 mediates strong inhibition of protein expression, *Mol. Cell. Biochem.*, vol. 446, n°1-2, p. 73-81.

Gutiérrez, L. G., Hernández-Morales, M., Núñez, L. et Villalobos, C., (2019), Inhibition of Polyamine Biosynthesis Reverses Ca2+ Channel Remodeling in Colon Cancer Cells, *Cancers (Basel)*, vol. 11, n°1, p. E83.

Hai, Y., Cao, W., Liu, G., Hong, S.-P., Elela, A. S., Klinck, R., … Xie, J., (2008), A G-tract element in apoptotic agents-induced alternative splicing., *Nucleic Acids Res.*, vol. 36, n°10, p. 3320-3331.

Halder, K., Benzler, M. et Hartig, J. S., (2012), Reporter assays for studying quadruplex nucleic acids, *Methods*, vol. 57, n°1, p. 115-121.

Halder, K. et Hartig, J. S., (2011), RNA quadruplexes, *Met. Ions Life Sci.*, vol. 9, p. 125-139.

Halder, K., Wieland, M. et Hartig, J. S., (2009), Predictable suppression of gene expression by 5′-UTR-based RNA quadruplexes, *Nucleic Acids Res.*, vol. 37, n°20, p. 6811-6817.

Hanahan, D. et Weinberg, R. A., (2011), Hallmarks of Cancer: The Next Generation, *Cell*, vol. 144, n°5, p. 646-674.

Hänsel-Hertsch, R., Antonio, M. D. et Balasubramanian, S., (2017), DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential, *Nat. Rev. Mol. Cell Biol.*, vol. 18, n°5, p. 279-284.

Hardin, C. C., Henderson, E., Watson, T. et Prosser, J. K., (1991), Monovalent cation induced structural transitions in telomeric DNAs : G-DNA folding intermediates, *Biochemistry*, vol. 30, n°18, p. 4460-4472.

Hardin, C. C., Perry, A. G. et White, K., (2000), Thermodynamic and kinetic characterization of the dissociation and assembly of quadruplex nucleic acids, *Biopolymers*, vol. 56, n°3, p. 147-194.

Hardin, C. C., Watson, T., Corregan, M. et Bailey, C., (1992), Cation-dependent transition between the quadruplex and Watson-Crick hairpin forms of d(CGCG3GCG), *Biochemistry*, vol. 31, n°3, p. 833-841.

Harris, L. M. et Merrick, C. J., (2015), G-quadruplexes in pathogens : a common route to virulence control ?, *PLoS Pathog.*, vol. 11, n°2, p. e1004562.

Hazel, P., Huppert, J., Balasubramanian, S. et Neidle, S., (2004), Loop-length-dependent folding of G-quadruplexes, *J. Am. Chem. Soc.*, vol. 126, n°50, p. 16405-16415.

Heddi, B., Martín-Pintado, N., Serimbetov, Z., Kari, T. M. A. et Phan, A. T., (2016), G-quadruplexes with (4n - 1) guanines in the G-tetrad core: formation of a G-triad·water complex and implication for small-molecule binding, *Nucleic Acids Res.*, vol. 44, n°2, p. 910-916.

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F. et Vandesompele, J., (2007), qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data, *Genome Biol.*, vol. 8, n°2, p. R19.

Henderson, A., Wu, Y., Huang, Y. C., Chavez, E. A., Platt, J., Johnson, F. B., … Lansdorp, P. M., (2014), Detection of G-quadruplex DNA in mammalian cells, *Nucleic Acids Res*, vol. 42, n°2, p. 860-869.

Herdy, B., Mayer, C., Varshney, D., Marsico, G., Murat, P., Taylor, C., … Balasubramanian, S., (2018a), Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a

novel interactor of cellular G-quadruplex containing transcripts, *Nucleic Acids Res.*, vol. 46, n°21, p. 11592-11604.

Herdy, B., Mayer, C., Varshney, D., Marsico, G., Murat, P., Taylor, C., … Balasubramanian, S., (2018b), Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a novel interactor of cellular G-quadruplex containing transcripts, *Nucleic Acids Res.*, vol. 46, n°21, p. 11592-11604.

Hinnebusch, A. G., Ivanov, I. P. et Sonenberg, N., (2016), Translational control by 5′-untranslated regions of eukaryotic mRNAs, *Science*, vol. 352, n°6292, p. 1413-1416.

H.J. Motulsky, (2014), *Intuitive Biostatistics*, Oxford University Press, 544 p.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. et Schuster, P., (1994), Fast folding and comparison of RNA secondary structures, *Monatsh. Chem.*, vol. 125, n°2, p. 167-188.

Hon, J., Martínek, T., Zendulka, J. et Lexa, M., (2017), pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R, *Bioinformatics*, vol. 33, n°21, p. 3373-3379.

Huang, D. W., Sherman, B. T. et Lempicki, R. A., (2009), Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res.*, vol. 37, n°1, p. 1-13.

Huang, H., Suslov, N. B., Li, N.-S., Shelke, S. A., Evans, M. E., Koldobskaya, Y., … Piccirilli, J. A., (2014), A G-quadruplex-containing RNA activates fluorescence in a GFP-like fluorophore, *Nat. Chem. Biol.*, vol. 10, n°8, p. 686-691.

Huang, H., Zhang, J., Harvey, S. E., Hu, X. et Cheng, C., (2017), RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF, *Genes Dev.*, vol. 31, n°22, p. 2296-2309.

Huang, L., Ashraf, S. et Lilley, D. M. J., (2019), The role of RNA structure in translational regulation by L7Ae protein in archaea, *RNA*, vol. 25, n°1, p. 60-69.

Huang, W., Liu, Z., Zhou, G., Ling, J., Tian, A. et Sun, N., (2016), Silencing Bag-1 gene via magnetic gold nanoparticle-delivered siRNA plasmid for colorectal cancer therapy in vivo and in vitro, *Tumour Biol.*, vol. 37, n°8, p. 10365-10374.

Huijbregts, L., Roze, C., Bonafe, G., Houang, M., Le Bouc, Y., Carel, J.-C., … de Roux, N., (2012), DNA polymorphisms of the KiSS1 3’ untranslated region interfere with the folding of a G-rich sequence into G-quadruplex, *Mol. Cell. Endocrinol.*, vol. 351, n°2, p. 239-248.

Huppert, J. L., (2008a), Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes, *Chem Soc Rev*, vol. 37, n°7, p. 1375-1384.

Huppert, J. L., (2008b), Hunting G-quadruplexes, *Biochimie*, vol. 90, n°8, p. 1140-1148.

Huppert, J. L. et Balasubramanian, S., (2005), Prevalence of quadruplexes in the human genome, *Nucleic Acids Res.*, vol. 33, n°9, p. 2908-2916.

Huppert, J. L., Bugaut, A., Kumari, S. et Balasubramanian, S., (2008), G-quadruplexes: the beginning and end of UTRs, *Nucleic Acids Res.*, vol. 36, n°19, p. 6260-6268.

Huppert, L. J. et Balasubramanian, S., (2007), G-quadruplexes in promoters throughout the human genome., *Nucleic Acids Res.*, vol. 35, n°2, p. 406-413.

Ibrahim, F., Maragkakis, M., Alexiou, P. et Mourelatos, Z., (2018), Ribothrypsis, a novel process of canonical mRNA decay, mediates ribosome-phased mRNA endonucleolysis, *Nat. Struct. Mol. Biol.*, vol. 25, n°4, p. 302-310.

Ingolia, N. T., Lareau, L. F. et Weissman, J. S., (2011), Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes, *Cell*, vol. 147, n°4, p. 789-802.

Ishiguro, A., Kimura, N., Watanabe, Y., Watanabe, S. et Ishihama, A., (2016), TDP-43 binds and transports G-quadruplex-containing mRNAs into neurites for local translation., *Genes Cells*, vol. 21, n°5, p. 466-81.

Ito, K., Go, S., Komiyama, M. et Xu, Y., (2011), Inhibition of translation by small RNA-stabilized mRNA structures in human cells, *J. Am. Chem. Soc.*, vol. 133, n°47, p. 19153-19159.

Ivanov, I. P., Firth, A. E., Michel, A. M., Atkins, J. F. et Baranov, P. V., (2011), Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences, *Nucleic Acids Res.*, vol. 39, n°10, p. 4220-4234.

Ivanov, P., O'Day, E., Emara, M. M., Wagner, G., Lieberman, J. et Anderson, P., (2014), G-quadruplex structures contribute to the neuroprotective effects of angiogenin-induced tRNA fragments., *Proc. Natl. Acad. Sci. U.S.A.*, vol. 111, n°51, p. 18201-18206.

Izbicka, E., Wheelhouse, R. T., Raymond, E., Davidson, K. K., Lawrence, R. A., Sun, D., … Hoff, D. D. V., (1999), Effects of Cationic Porphyrins as G-Quadruplex Interactive Agents in Human Tumor Cells, *Cancer Res.*, vol. 59, n°3, p. 639-644.

Jacobs, G. H., Chen, A., Stevens, S. G., Stockwell, P. A., Black, M. A., Tate, W. P. et Brown, C. M., (2009), Transterm : a database to aid the analysis of regulatory sequences in mRNAs, *Nucleic Acids Res.*, vol. 37, n°Database issue, p. D72-76.

Jayaraj, G. G., Pandey, S., Scaria, V. et Maiti, S., (2012), Potential G-quadruplexes in the human long non-coding transcriptome, *RNA Biol.*, vol. 9, n°1, p. 81-89.

Jiang, J., Chan, H., Cash, D. D., Miracco, E. J., Ogorzalek Loo, R. R., Upton, H. E., … Feigon, J., (2015), Structure of Tetrahymena telomerase reveals previously unknown subunits, functions, and interactions, *Science*, vol. 350, n°6260, p. aab4070.

Joachimi, A., Benz, A. et Hartig, J. S., (2009), A comparison of DNA and RNA quadruplex structures and stabilities, *Bioorg. Med. Chem.*, vol. 17, n°19, p. 6811-6815.

Jodoin, R., Bauer, L., Garant, J.-M., Mahdi Laaref, A., Phaneuf, F. et Perreault, J.-P., (2014), The folding of 5'-UTR human G-quadruplexes possessing a long central loop., *RNA*, vol. 20, p. 1129-1141.

Jodoin, R. et Perreault, J.-P., (2018), G-quadruplexes formation in the 5'UTRs of mRNAs associated with colorectal cancer pathways, *PloS One*, vol. 13, n°12, p. e0208363.

Johnstone, T. G., Bazzini, A. A. et Giraldez, A. J., (2016), Upstream ORFs are prevalent translational repressors in vertebrates, *EMBO J.*, vol. 35, n°7, p. 706-723.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. et Morishima, K., (2017), KEGG : new perspectives on genomes, pathways, diseases and drugs, *Nucleic Acids Res.*, vol. 45, n°D1, p. D353-D361.

Karabiber, F., McGinnis, J. L., Favorov, O. V. et Weeks, K. M., (2013), QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis, *RNA*, vol. 19, n°1, p. 63-73.

Karsisiotis, A. I., O'Kane, C. et Webba da Silva, M., (2013), DNA quadruplex folding formalism--a tutorial on quadruplex topologies, *Methods*, vol. 64, n°1, p. 28-35.

Katsuda, Y., Sato, S.-I., Asano, L., Morimura, Y., Furuta, T., Sugiyama, H., … Uesugi, M., (2016), A Small Molecule That Represses Translation of G-Quadruplex-Containing mRNA, *J. Am. Chem. Soc.*, vol. 138, n°29, p. 9037-9040.

Kazemier, H. G., Paeschke, K. et Lansdorp, P. M., (2017), Guanine quadruplex monoclonal antibody 1H6 cross-reacts with restrained thymidine-rich single stranded DNA, *Nucleic Acids Res.*, vol. 45, n°10, p. 5913-5919.

Kenny, P. J., Zhou, H., Kim, M., Skariah, G., Khetani, R. S., Drnevich, J., … Ceman, S., (2014), MOV10 and FMRP Regulate AGO2 Association with MicroRNA Recognition Elements, *Cell Rep.*, vol. 9, n°5, p. 1729-1741.

Khateb, S., Weisman-Shomer, P., Hershco-Shani, I., Ludwig, L. A. et Fry, M., (2007), The tetraplex (CGG)n destabilizing proteins hnRNP A2 and CBF-A enhance the in vivo translation of fragile X premutation mRNA., *Nucleic Acids Res.*, vol. 35, n°17, p. 5775-5788.

Kikin, O., D'Antonio, L. et Bagga, P. S., (2006), QGRS Mapper : a web-based server for predicting G-quadruplexes in nucleotide sequences, *Nucleic Acids Res.*, vol. 34, n°Web Server issue, p. W676-682.

Kikin, O., Zappala, Z., D'Antonio, L. et Bagga, P. S., (2008), GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs, *Nucleic Acids Res.*, vol. 36, n°Database issue, p. D141-148.

Kikuchi, R., Noguchi, T., Takeno, S., Funada, Y., Moriyama, H. et Uchida, Y., (2002), Nuclear BAG-1 expression reflects malignant potential in colorectal carcinomas, *Br. J. Cancer*, vol. 87, n°10, p. 1136-1139.

Koirala, D., Ghimire, C., Bohrer, C., Sannohe, Y., Sugiyama, H. et Mao, H., (2013), Long-loop G-quadruplexes are misfolded population minorities with fast transition kinetics in human telomeric sequences, *J. Am. Chem. Soc.*, vol. 135, n°6, p. 2235-2241.

Kosman, J. et Juskowiak, B., (2016), Hemin/G-quadruplex structure and activity alteration induced by magnesium cations, *Int. J. Biol. Macromol.*, vol. 85, p. 555-564.

Koukouraki, P. et Doxakis, E., (2016), Constitutive translation of human α-synuclein is mediated by the 5'-untranslated region, *Open Biol*, vol. 6, n°4, p. 160022.

Kozak, M., (1987), At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells, *J. Mol. Biol.*, vol. 196, n°4, p. 947-950.

Kozak, M., (1989), Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs., *Mol. Cell. Biol.*, vol. 9, n°11, p. 5134-5142.

Kumar, N. et Maiti, S., (2005), The effect of osmolytes and small molecule on Quadruplex-WC duplex equilibrium: a fluorescence resonance energy transfer study, *Nucleic Acids Res.*, vol. 33, n°21, p. 6723-6732.

Kumar, N. et Maiti, S., (2008), A thermodynamic overview of naturally occurring intramolecular DNA quadruplexes, *Nucleic Acids Res.*, vol. 36, n°17, p. 5610-5622.

Kumar, N., Sahoo, B., Varun, K. A. S., Maiti, S. et Maiti, S., (2008), Effect of loop length variation on quadruplex-Watson Crick duplex competition, *Nucleic Acids Res.*, vol. 36, n°13, p. 4433-4442.

Kumari, R., Nambiar, M., Shanbagh, S. et Raghavan, S. C., (2015), Detection of G-Quadruplex DNA Using Primer Extension as a Tool, *PLoS One*, vol. 10, n°3, p. e0119722.

Kumari, S., Bugaut, A. et Balasubramanian, S., (2008), Position and stability are determining factors for translation repression by an RNA G-quadruplex-forming sequence within the 5' UTR of the NRAS proto-oncogene., *Biochemistry*, vol. 47, n°48, p. 12664-12669.

Kumari, S., Bugaut, A., Huppert, L. J. et Balasubramanian, S., (2007), An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation., *Nat. Chem. Biol.*, vol. 3, n°4, p. 218-221.

Kuo, H.-J. M., Wang, Z.-F., Tseng, T.-Y., Li, M.-H., Hsu, D. S.-T., Lin, J.-J. et Chang, T.-C., (2015a), Conformational transition of a hairpin structure to G-quadruplex within the WNT1 gene promoter., *J. Am. Chem. Soc.*, vol. 137, n°1, p. 210-218.

Kuo, M. H.-J., Wang, Z.-F., Tseng, T.-Y., Li, M.-H., Hsu, S.-T. D., Lin, J.-J. et Chang, T.-C., (2015b), Conformational Transition of a Hairpin Structure to G-Quadruplex within the WNT1 Gene Promoter, *J. Am. Chem. Soc.*, vol. 137, n°1, p. 210-218.

Kwok, C. K. et Balasubramanian, S., (2015), Targeted Detection of G-Quadruplexes in Cellular RNAs, *Angew. Chem. Int. Ed.*, vol. 54, n°23, p. 6751-6754.

Kwok, C. K., Marsico, G. et Balasubramanian, S., (2018), Detecting RNA G-Quadruplexes (rG4s) in the Transcriptome, *Cold Spring Harb Perspect Biol*, vol. 10, n°7, p. a032284.

Kwok, C. K., Marsico, G., Sahakyan, A. B., Chambers, V. S. et Balasubramanian, S., (2016a), rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome, *Nat. Methods*, vol. 13, n°10, p. 841-844.

Kwok, C. K., Sahakyan, A. B. et Balasubramanian, S., (2016b), Structural Analysis using SHALiPE to Reveal RNA G-Quadruplex Formation in Human Precursor MicroRNA, *Angew. Chem., Int. Ed. Engl.*, vol. 55, n°31, p. 8958-8961.

Kwok, C. K., Sherlock, M. E. et Bevilacqua, P. C., (2013), Effect of loop sequence and loop length on the intrinsic fluorescence of G-quadruplexes, *Biochemistry*, vol. 52, n°18, p. 3019-3021.

Kwok, K. C., Ding, Y., Shahid, S., Assmann, M. S. et Bevilacqua, C. P., (2015), A stable RNA G-quadruplex within the 5'-UTR of Arabidopsis thaliana ATR mRNA inhibits translation., *Biochem. J.*, vol. 467, n°1, p. 91-102.

Laederach, A., Das, R., Vicens, Q., Pearlman, S. M., Brenowitz, M., Herschlag, D. et Altman, R. B., (2008), Semiautomated and rapid quantification of nucleic acid footprinting and structure mapping experiments, *Nat. Protoc.*, vol. 3, n°9, p. 1395-1401.

Lago, S., Tosoni, E., Nadai, M., Palumbo, M. et Richter, S. N., (2017), The cellular protein nucleolin preferentially binds long-looped G-quadruplex nucleic acids, *Biochim. Biophys. Acta Gen. Subj.*, vol. 1861, n°5 Pt B, p. 1371-1381.

Laguerre, A., Hukezalie, K., Winckler, P., Katranji, F., Chanteloup, G., Pirrotta, M., … Monchaud, D., (2015), Visualization of RNA-Quadruplexes in Live Cells, *J. Am. Chem. Soc.*, vol. 137, n°26, p. 8521-8525.

Laguerre, A., Wong, J. M. Y. et Monchaud, D., (2016), Direct visualization of both DNA and RNA quadruplexes in human cells *via* an uncommon spectroscopic method, *Sci. Rep.*, vol. 6, p. 32141.

Lai, D., Proctor, J. R., Zhu, J. Y. A. et Meyer, I. M., (2012), R- chie : a web server and R package for visualizing RNA secondary structures, *Nucleic Acids Res.*, vol. 40, n°12, p. e95-e95.

Lai, H., Xiao, Y., Yan, S., Tian, F., Zhong, C., Liu, Y., … Zhou, X., (2014), Symmetric cyanovinyl-pyridinium triphenylamine: a novel fluorescent switch-on probe for an antiparallel G-quadruplex, *Analyst*, vol. 139, n°8, p. 1834-1838.

Lai, M.-C., Chang, C.-M. et Sun, H. S., (2016), Hypoxia Induces Autophagy through Translational Up-Regulation of Lysosomal Proteins in Human Colon Cancer Cells, *PLOS ONE*, vol. 11, n°4, p. e0153627.

Lamas, M., Monaco, L., Zazopoulos, E., Lalli, E., Tamai, K., Penna, L., … Sassone-Corsi, P., (1996), CREM: a master-switch in the transcriptional response to cAMP, *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, vol. 351, n°1339, p. 561-567.

Lammich, S., Kamp, F., Wagner, J., Nuscher, B., Zilow, S., Ludwig, A.-K., … Haass, C., (2011), Translational repression of the disintegrin and metalloprotease ADAM10 by a stable G-quadruplex secondary structure in its 5'-untranslated region., *J. Biol. Chem.*, vol. 286, n°52, p. 45063-45072.

Lane, A. N., Chaires, J. B., Gray, R. D. et Trent, J. O., (2008), Stability and kinetics of G-quadruplex structures, *Nucleic Acids Res.*, vol. 36, n°17, p. 5482-5515.

Lattmann, S., Stadler, M. B., Vaughn, J. P., Akman, S. A. et Nagamine, Y., (2011), The DEAH-box RNA helicase RHAU binds an intramolecular RNA G-quadruplex in TERC and associates with telomerase holoenzyme, *Nucleic Acids Res.*, vol. 39, n°21, p. 9390-9404.

Lavezzo, E., Berselli, M., Frasson, I., Perrone, R., Palù, G., Brazzale, A. R., … Toppo, S., (2018), G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide, *PLoS Comput. Biol.*, vol. 14, n°12, p. e1006675.

Lee, S. C., Zhang, J., Strom, J., Yang, D., Dinh, T. N., Kappeler, K. et Chen, Q. M., (2017), G-Quadruplex in the NRF2 mRNA 5' Untranslated Region Regulates De Novo NRF2 Protein Translation under Oxidative Stress, *Mol. Cell. Biol.*, vol. 37, n°1, p. e00122-16.

Lee, S., Kim, H., Tian, S., Lee, T., Yoon, S. et Das, R., (2015), Automated band annotation for RNA structure probing experiments with numerous capillary electrophoresis profiles, *Bioinformatics*, vol. 31, n°17, p. 2808-2815.

Leeder, W.-M., Hummel, N. F. C. et Göringer, H. U., (2016), Multiple G-quartet structures in pre-edited mRNAs suggest evolutionary driving force for RNA editing in trypanosomes, *Sci. Rep.*, vol. 6, p. 29810.

Leppek, K., Das, R. et Barna, M., (2018), Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them, *Nat. Rev. Mol. Cell Biol.*, vol. 19, n°3, p. 158-174.

Li, Q., Xiang, J.-F., Yang, Q.-F., Sun, H.-X., Guan, A.-J. et Tang, Y.-L., (2013), G4LDB : a database for discovering and studying G-quadruplex ligands, *Nucleic Acids Res.*, vol. 41, n°Database issue, p. D1115-1123.

Li, X., Zheng, K., Zhang, J., Liu, H., He, Y., Yuan, B., … Tan, Z., (2015), Guanine-vacancy-bearing G-quadruplexes responsive to guanine derivatives, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, n°47, p. 14581-14586.

Li, Y. et Breaker, R. R., (1999), Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2‘-Hydroxyl Group, *J. Am. Chem. Soc.*, vol. 121, n°23, p. 5364-5372.

Lightfoot, H. L., Hagen, T., Cléry, A., Allain, F. H.-T. et Hall, J., (2018), Control of the polyamine biosynthesis pathway by G2-quadruplexes, *ELife*, vol. 7, p. e36362.

Lipps, H. J. et Rhodes, D., (2009), G-quadruplex structures: in vivo evidence and function, *Trends Cell Biol.*, vol. 19, n°8, p. 414-422.

Liu, H., Matsugami, A., Katahira, M. et Uesugi, S., (2002), A dimeric RNA quadruplex architecture comprised of two G:G(:A):G:G(:A) hexads, G:G:G:G tetrads and UUUU loops, *J. Mol. Biol.*, vol. 322, n°5, p. 955-970.

Liu, X. et Xu, Y., (2018), HnRNPA1 Specifically Recognizes the Base of Nucleotide at the Loop of RNA G-Quadruplex, *Molecules*, vol. 23, n°1, .

Lorenz, R., Bernhart, S. H., Externbrink, F., Qin, J., Höner zu Siederdissen, C., Amman, F., … Stadler, P. F., (2012), RNA Folding Algorithms with G-Quadruplexes, In : M. C. DE SOUTO & M. G. KANN (éd.), *Advances in Bioinformatics and Computational Biology*, Springer Berlin Heidelberg, p. 49-60.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F. et Hofacker, I. L., (2011), ViennaRNA Package 2.0, *Algorithms Mol. Biol.*, vol. 6, p. 26.

Lorenz, R., Bernhart, S. H., Qin, J., Höner zu Siederdissen, C., Tanzer, A., Amman, F., … Stadler, P. F., (2013), 2D meets 4G : G-quadruplexes in RNA secondary structure prediction, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 10, n°4, p. 832-844.

Low, J. T. et Weeks, K. M., (2010), SHAPE-directed RNA secondary structure prediction, *Methods*, vol. 52, n°2, p. 150-158.

Lüders, J., Demand, J. et Höhfeld, J., (2000), The Ubiquitin-related BAG-1 Provides a Link between the Molecular Chaperones Hsc70/Hsp70 and the Proteasome, *J. Biol. Chem.*, vol. 275, n°7, p. 4613-4617.

Lung Chan, K., Peng, B., I. Umar, M., Chan, C.-Y., B. Sahakyan, A., N. Le, M. T. et Kit Kwok, C., (2018), Structural analysis reveals the formation and role of RNA G-quadruplex structures in human mature microRNAs, *Chem. Commun.*, vol. 54, n°77, p. 10878-10881.

Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A. et Sampath, R., (2001), RNAMotif, an RNA secondary structure definition and search algorithm, *Nucleic Acids Res.*, vol. 29, n°22, p. 4724-4735.

Mailler, E., Paillart, J.-C., Marquet, R., Smyth, R. P. et Vivet-Boudou, V., (2018), The evolution of RNA structural probing methods: From gels to next-generation sequencing, *Wiley Interdiscip. Rev. RNA*, p. e1518.

Maizels, N., (2015), G4-associated human diseases, *EMBO Reports*, vol. 16, n°8, p. 910-922.

Malgowska, M., Czajczynska, K., Gudanis, D., Tworak, A. et Gdaniec, Z., (2016), Overview of the RNA G-quadruplex structures, *Acta Biochim. Pol.*, vol. 63, n°4, p. 609-621.

Malgowska, M., Gudanis, D., Kierzek, R., Wyszko, E., Gabelica, V. et Gdaniec, Z., (2014), Distinctive structural motifs of RNA G-quadruplexes composed of AGG, CGG and UGG trinucleotide repeats., *Nucleic Acids Res.*, vol. 42, n°15, p. 10196-10207.

Manna, S. et Srivatsan, S. G., (2018), Fluorescence-based tools to probe G-quadruplexes in cell-free and cellular environments, *RSC Adv.*, vol. 8, n°45, p. 25673-25694.

Mao, S.-Q., Ghanbarian, A. T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M., … Balasubramanian, S., (2018), DNA G-quadruplex structures mold the DNA methylome, *Nat. Struct. Mol. Biol.*, vol. 25, n°10, p. 951-957.

Marcel, V., Tran, T. P. L., Sagne, C., Martel-Planche, G., Vaslin, L., Teulade-Fichou, M.-P., … Van Dyck, E., (2011), G-quadruplex structures in TP53 intron 3 : role in alternative splicing and in production of p53 mRNA isoforms., *Carcinogenesis*, vol. 32, n°3, p. 271-278.

Marsico, G., Chambers, V. S., Sahakyan, A. B., McCauley, P., Boutell, J. M., Di Antonio, M. et Balasubramanian, S., (2019), Whole genome experimental maps of DNA G-quadruplexes in multiple species, *Nucleic Acids Res*.

Martadinata, H., Heddi, B., Lim, W. K. et Phan, T. A., (2011), Structure of long human telomeric RNA (TERRA) : G-quadruplexes formed by four and eight UUAGGG repeats are stable building blocks., *Biochemistry*, vol. 50, n°29, p. 6455-6461.

Martadinata, H. et Phan, A. T., (2014), Formation of a stacked dimeric G-quadruplex containing bulges by the 5'-terminal region of human telomerase RNA (hTERC), *Biochemistry*, vol. 53, n°10, p. 1595-1600.

Martino, L., Virno, A., Pagano, B., Virgilio, A., Di Micco, S., Galeone, A., … Randazzo, A., (2007), Structural and Thermodynamic Studies of the Interaction of Distamycin A with the Parallel Quadruplex Structure [d(TGGGGT)]4, *J. Am. Chem. Soc.*, vol. 129, n°51, p. 16048-16056.

Matsumura, K., Kawasaki, Y., Miyamoto, M., Kamoshida, Y., Nakamura, J., Negishi, L., … Akiyama, T., (2017), The novel G-quadruplex-containing long non-coding RNA GSEC antagonizes DHX36 and modulates colon cancer cell migration, *Oncogene*, vol. 36, n°9, p. 1191-1199.

McAninch, D. S., Heinaman, A. M., Lang, C. N., Moss, K. R., Bassell, G. J., Rita Mihailescu, M. et Evans, T. L., (2017), Fragile X mental retardation protein recognizes a G quadruplex

structure within the survival motor neuron domain containing 1 mRNA 5'-UTR, *Mol. Biosyst.*, vol. 13, n°8, p. 1448-1457.

McLuckie, K. I. E., Di Antonio, M., Zecchini, H., Xian, J., Caldas, C., Krippendorff, B.-F., … Balasubramanian, S., (2013), G-quadruplex DNA as a molecular target for induced synthetic lethality in cancer cells, *J. Am. Chem. Soc.*, vol. 135, n°26, p. 9640-9643.

McRae, E. K. S., Booy, E. P., Moya-Torres, A., Ezzati, P., Stetefeld, J. et McKenna, S. A., (2017), Human DDX21 binds and unwinds RNA guanine quadruplexes, *Nucleic Acids Res.*, vol. 45, n°11, p. 6656-6668.

Melko, M., Douguet, D., Bensaid, M., Zongaro, S., Verheggen, C., Gecz, J. et Bardoni, B., (2011), Functional characterization of the AFF (AF4/FMR2) family of RNA-binding proteins: insights into the molecular pathology of FRAXE intellectual disability, *Hum. Mol. Genet.*, vol. 20, n°10, p. 1873-1885.

Mendoza, O., Bourdoncle, A., Boulé, J.-B., Brosh, M. R. et Mergny, J.-L., (2016), G-quadruplexes and helicases., *Nucleic Acids Res.*, vol. 44, n°5, p. 1989-2006.

Mergny, J.-L., De Cian, A., Ghelab, A., Saccà, B. et Lacroix, L., (2005), Kinetics of tetramolecular quadruplexes, *Nucleic Acids Res.*, vol. 33, n°1, p. 81-94.

Mergny, J.-L. et Lacroix, L., (2009), UV Melting of G-Quadruplexes, *Curr. Protoc. Nucleic Acid Chem.*, vol. 37, n°1, p. 17.1.1-17.1.15.

Meyers, R. A., (2004), *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, Wiley-Blackwell, 716 p.

Michel, A. M., Fox, G., M. Kiran, A., De Bo, C., O'Connor, P. B. F., Heaphy, S. M., … Baranov, P. V., (2014), GWIPS-viz: development of a ribo-seq genome browser, *Nucleic Acids Res.*, vol. 42, n°D1, p. D859-D864.

Miglietta, G., Cogoi, S., Marinello, J., Capranico, G., Tikhomirov, A. S., Shchekotikhin, A. et Xodo, L. E., (2017), RNA G-Quadruplexes in Kirsten Ras (KRAS) Oncogene as Targets for Small Molecules Inhibiting Translation, *J. Med. Chem.*, vol. 60, n°23, p. 9448-9461.

Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P. J., … Pesole, G., (2005), UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs, *Nucleic Acids Res.*, vol. 33, n°Database issue, p. D141-146.

Millevoi, S., Moine, H. et Vagner, S., (2012), G-quadruplexes in RNA biology, *Wiley Interdiscip. Rev. : RNA*, vol. 3, n°4, p. 495-507.

Mirihana Arachchilage, G., Morris, M. J. et Basu, S., (2014), A library screening approach identifies naturally occurring RNA sequences for a G-quadruplex binding ligand, *Chem. Commun. (Camb.)*, vol. 50, n°10, p. 1250-1252.

Miyoshi, D., Fujimoto, T. et Sugimoto, N., (2013), Molecular Crowding and Hydration Regulating of G-Quadruplex Formation, In : J. B. CHAIRES & D. GRAVES (éd.), *Quadruplex Nucleic Acids*, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 87-110.

Miyoshi, D., Matsumura, S., Nakano, S.-I. et Sugimoto, N., (2004), Duplex dissociation of telomere DNAs induced by molecular crowding, *J. Am. Chem. Soc.*, vol. 126, n°1, p. 165-169.

Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J.-D., Fernandez, J. P., Mis, E. K., Khokha, M. K. et Giraldez, A. J., (2015), CRISPRscan : designing highly efficient sgRNAs for CRISPR/Cas9 targeting in vivo, *Nat Methods*, vol. 12, n°10, p. 982-988.

Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B. et Kinzler, K. W., (1997), Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC, *Science*, vol. 275, n°5307, p. 1787-1790.

Morris, J. M. et Basu, S., (2009), An unusually stable G-quadruplex within the 5'-UTR of the MT3 matrix metalloproteinase mRNA represses translation in eukaryotic cells., *Biochemistry*, vol. 48, n°23, p. 5313-5319.

Morris, J. M., Negishi, Y., Pazsint, C., Schonhoft, D. J. et Basu, S., (2010), An RNA G-quadruplex is essential for cap-independent translation initiation in human VEGF IRES., *J. Am. Chem. Soc.*, vol. 132, n°50, p. 17831-17839.

Morris, M. J., Wingate, K. L., Silwal, J., Leeper, T. C. et Basu, S., (2012), The porphyrin TmPyP4 unfolds the extremely stable G-quadruplex in MT3-MMP mRNA and alleviates its repressive effect to enhance translation in eukaryotic cells, *Nucleic Acids Res.*, vol. 40, n°9, p. 4137-4145.

Mortimer, S. A., Kidwell, M. A. et Doudna, J. A., (2014), Insights into RNA structure and function from genome-wide studies, *Nature Reviews Genetics*, vol. 15, n°7, p. 469-479.

Mukohyama, J., Shimono, Y., Minami, H., Kakeji, Y. et Suzuki, A., (2017), Roles of microRNAs and RNA-Binding Proteins in the Regulation of Colorectal Cancer Stem Cells, *Cancers (Basel)*, vol. 9, n°10, .

Mukundan, V. T. et Phan, A. T., (2013), Bulges in G-quadruplexes : broadening the definition of G-quadruplex-forming sequences, *J. Am. Chem. Soc.*, vol. 135, n°13, p. 5017-5028.

Mullen, M. A., Olson, K. J., Dallaire, P., Major, F., Assmann, S. M. et Bevilacqua, P. C., (2010), RNA G-Quadruplexes in the model plant species Arabidopsis thaliana: prevalence and possible functional roles, *Nucleic Acids Res.*, vol. 38, n°22, p. 8149-8163.

Murat, P., Marsico, G., Herdy, B., Ghanbarian, A., Portella, G. et Balasubramanian, S., (2018), RNA G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs, *Genome Biol.*, vol. 19, n°1, p. 229.

Murat, P., Zhong, J., Lekieffre, L., Cowieson, N. P., Clancy, J. L., Preiss, T., … Tellam, J., (2014), G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation, *Nat. Chem. Biol.*, vol. 10, n°5, p. 358-364.

Nakano, S., Miyoshi, D. et Sugimoto, N., (2014), Effects of Molecular Crowding on the Structures, Interactions, and Functions of Nucleic Acids, *Chem. Rev.*, vol. 114, n°5, p. 2733-2758.

Neidle, S., (2012), *Therapeutic applications of quadruplex nucleic acids*, Amsterdam : Elsevier/Academic Press, 196 p.

Neidle, S. et Parkinson, G. N., (2008), Quadruplex DNA crystal structures and drug design, *Biochimie*, vol. 90, n°8, p. 1184-1196.

Nicoludis, J. M., Barrett, S. P., Mergny, J.-L. et Yatsunyk, L. A., (2012), Interaction of human telomeric DNA with N- methyl mesoporphyrin IX, *Nucleic Acids Res.*, vol. 40, n°12, p. 5432-5447.

Nieradka, A., Ufer, C., Thiadens, K., Grech, G., Horos, R., van Coevorden-Hameete, M., … von Lindern, M., (2014), Grsf1-Induced Translation of the SNARE Protein Use1 Is Required for Expansion of the Erythroid Compartment, *PLoS One*, vol. 9, n°9, p. e104631.

Noderer, W. L., Flockhart, R. J., Bhaduri, A., Arce, A. J. D. de, Zhang, J., Khavari, P. A. et Wang, C. L., (2014), Quantitative analysis of mammalian translation initiation sites by FACS-seq, *Mol. Syst. Biol.*, vol. 10, n°8, p. 748.

O'Day, E., Le, M. T. N., Imai, S., Tan, S. M., Kirchner, R., Arthanari, H., … Lieberman, J., (2015), An RNA-binding Protein, Lin28, Recognizes and Remodels G-quartets in the MicroRNAs (miRNAs) and mRNAs It Regulates, *J. Biol. Chem.*, vol. 290, n°29, p. 17909-17922.

Olsen, C. M., Lee, H.-T. et Marky, L. A., (2009), Unfolding thermodynamics of intramolecular G-quadruplexes : base sequence contributions of the loops, *J. Phys. Chem. B*, vol. 113, n°9, p. 2587-2595.

Olsen, C. M. et Marky, L. A., (2009), Energetic and hydration contributions of the removal of methyl groups from thymine to form uracil in G-quadruplexes, *J. Phys. Chem. B*, vol. 113, n°1, p. 9-11.

Olsthoorn, R. C. L., (2014), G-quadruplexes within prion mRNA: the missing link in prion disease?, *Nucleic Acids Res.*, vol. 42, n°14, p. 9327-9333.

Packham, G., Brimmell, M. et Cleveland, L. J., (1997), Mammalian cells express two differently localized Bag-1 isoforms generated by alternative translation initiation, *Biochem. J.*, vol. 328, n°3, p. 807-813.

Pandey, S., Agarwala, P., Jayaraj, G. G., Gargallo, R. et Maiti, S., (2015), The RNA Stem-Loop to G-Quadruplex Equilibrium Controls Mature MicroRNA Production inside the Cell., *Biochemistry*, vol. 54, n°48, p. 7067-7078.

Pandey, S., Agarwala, P. et Maiti, S., (2013), Effect of loops and G-quartets on the stability of RNA G-quadruplexes, *J. Phys. Chem. B*, vol. 117, n°23, p. 6896-6905.

Papadakis, E. S., Barker, C. R., Syed, H., Reeves, T., Schwaiger, S., Stuppner, H., … Cutress, R. I., (2016), The Bag-1 inhibitor, Thio-2, reverses an atypical 3D morphology driven by Bag-1L overexpression in a MCF-10A model of ductal carcinoma in situ, *Oncogenesis*, vol. 5, n°4, p. e215.

Paramasivan, S., Rujan, I. et Bolton, P. H., (2007), Circular dichroism of quadruplex DNAs : applications to structure, cation effects and ligand binding, *Methods*, vol. 43, n°4, p. 324-331.

Parrotta, L., Ortuso, F., Moraca, F., Rocca, R., Costa, G., Alcaro, S. et Artese, A., (2014), Targeting unimolecular G-quadruplex nucleic acids : a new paradigm for the drug discovery?, *Expert Opin. Drug Discovery*, vol. 9, n°10, p. 1167-1187.

Patel, D. J., Phan, A. T. et Kuryavyi, V., (2007), Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics, *Nucleic Acids Res.*, vol. 35, n°22, p. 7429-7455.

Perriaud, L., Marcel, V., Sagne, C., Favaudon, V., Guédin, A., De Rache, A., … Hall, J., (2014), Impact of G-quadruplex structures and intronic polymorphisms rs17878362 and rs1642785 on basal and ionizing radiation-induced expression of alternative p53 transcripts., *Carcinogenesis*, vol. 35, n°12, p. 2706-2715.

Phan, A. T., Kuryavyi, V., Darnell, J. C., Serganov, A., Majumdar, A., Ilin, S., … Patel, D. J., (2011), Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction, *Nat. Struct. Mol. Biol.*, vol. 18, n°7, p. 796-804.

Pickering, B. M., Mitchell, S. A., Spriggs, K. A., Stoneley, M. et Willis, A. E., (2004), Bag-1 Internal Ribosome Entry Segment Activity Is Promoted by Structural Changes Mediated by Poly(rC) Binding Protein 1 and Recruitment of Polypyrimidine Tract Binding Protein 1, *Mol. Cell. Biol.*, vol. 24, n°12, p. 5595-5605.

Prioleau, M.-N., (2017), G-Quadruplexes and DNA Replication Origins, *Adv. Exp. Med. Biol.*, vol. 1042, p. 273-286.

Provenzani, A., Fronza, R., Loreni, F., Pascale, A., Amadio, M. et Quattrone, A., (2006), Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis, *Carcinogenesis*, vol. 27, n°7, p. 1323-1333.

Rachwal, P. A., Brown, T. et Fox, K. R., (2007a), Sequence effects of single base loops in intramolecular quadruplex DNA, *FEBS Lett.*, vol. 581, n°8, p. 1657-1660.

Rachwal, P. A., Findlow, I. S., Werner, J. M., Brown, T. et Fox, K. R., (2007b), Intramolecular DNA quadruplexes with different arrangements of short and long loops, *Nucleic Acids Res.*, vol. 35, n°12, p. 4214-4222.

Randazzo, A., Spada, G. P., Silva, M. W. da, Spada, G. P. et Randazzo, A., (2012), Circular Dichroism of Quadruplex Structures, In : *Quadruplex Nucleic Acids*, Springer, Berlin, Heidelberg, p. 67-86.

Rawal, P., Kummarasetti, R. V. B., Ravindran, J., Kumar, N., Halder, K., Sharma, R., … Chowdhury, S., (2006), Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation., *Genome Res.*, vol. 16, n°5, p. 644-655.

Read, M., Harrison, R. J., Romagnoli, B., Tanious, F. A., Gowan, S. H., Reszka, A. P., … Neidle, S., (2001), Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, n°9, p. 4844-4849.

Regulski, E. E. et Breaker, R. R., (2008), In-line probing analysis of riboswitches, *Methods Mol. Biol.*, vol. 419, p. 53-67.

Renaud de la Faverie, A., Guédin, A., Bedrat, A., Yatsunyk, L. A. et Mergny, J.-L., (2014), Thioflavin T as a fluorescence light-up probe for G4 formation, *Nucleic Acids Res.*, vol. 42, n°8, p. e65.

Reuter, J. S. et Mathews, D. H., (2010), RNAstructure: software for RNA secondary structure prediction and analysis, *BMC Bioinf.*, vol. 11, p. 129.

Rihan, K., Antoine, E., Maurin, T., Bardoni, B., Bordonné, R., Soret, J. et Rage, F., (2017), A new cis-acting motif is required for the axonal SMN-dependent Anxa2 mRNA localization, *RNA*, vol. 23, n°6, p. 899-909.

Risitano, A. et Fox, K. R., (2004), Influence of loop size on the stability of intramolecular DNA quadruplexes, *Nucleic Acids Res.*, vol. 32, n°8, p. 2598-2606.

Robichaud, N., Sonenberg, N., Ruggero, D. et Schneider, R. J., (2018), Translational Control in Cancer, *Cold Spring Harbor Perspect. Biol.*, p. a032896.

Rocca, R., Talarico, C., Moraca, F., Costa, G., Romeo, I., Ortuso, F., … Artese, A., (2017), Molecular recognition of a carboxy pyridostatin toward G-quadruplex structures: Why does it prefer RNA?, *Chem. Biol. Drug. Des.*, vol. 90, n°5, p. 919-925.

Rodriguez, R., Müller, S., Yeoman, J. A., Trentesaux, C., Riou, J.-F. et Balasubramanian, S., (2008), A Novel Small Molecule That Alters Shelterin Integrity and Triggers a DNA-Damage Response at Telomeres, *J. Am. Chem. Soc.*, vol. 130, n°47, p. 15758-15759.

Rouleau, S. G., Beaudoin, J.-D., Bisaillon, M. et Perreault, J.-P., (2015), Small antisense oligonucleotides against G-quadruplexes : specific mRNA translational switches, *Nucleic Acids Res.*, vol. 43, n°1, p. 595-606.

Rouleau, S. G., Garant, J.-M., Bolduc, F., Bisaillon, M. et Perreault, J.-P., (2018), G-Quadruplexes influence pri-microRNA processing, *RNA Biol*, vol. 15, n°2, p. 198-206.

Rouleau, S., Glouzon, J.-P. S., Brumwell, A., Bisaillon, M. et Perreault, J.-P., (2017a), 3' UTR G-quadruplexes regulate miRNA binding, *RNA*, vol. 23, n°8, p. 1172-1179.

Rouleau, S., Jodoin, R., Garant, J.-M. et Perreault, J.-P., (2017b), RNA G-Quadruplexes as Key Motifs of the Transcriptome, In : *SpringerLink*, Springer, Berlin, Heidelberg, p. 1-20.

Sabharwal, N. C., Savikhin, V., Turek-Herman, J. R., Nicoludis, J. M., Szalai, V. A. et Yatsunyk, L. A., (2014), N-methylmesoporphyrin IX fluorescence as a reporter of strand orientation in guanine quadruplexes, *FEBS J.*, vol. 281, n°7, p. 1726-1737.

Sabouri, N., Capra, J. A. et Zakian, V. A., (2014), The essential Schizosaccharomyces pombe Pfh1 DNA helicase promotes fork movement past G-quadruplex motifs to prevent DNA damage, *BMC Biol.*, vol. 12, p. 101.

Saccà, B., Lacroix, L. et Mergny, J.-L., (2005), The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides, *Nucleic Acids Res.*, vol. 33, n°4, p. 1182-1192.

Sagliocco, F. A., Vega Laso, M. R., Zhu, D., Tuite, M. F., McCarthy, J. E. et Brown, A. J., (1993), The influence of 5'-secondary structures upon ribosome binding to mRNA during translation in yeast, *J. Biol. Chem.*, vol. 268, n°35, p. 26522-26530.

Sahakyan, A. B., Chambers, V. S., Marsico, G., Santner, T., Di Antonio, M. et Balasubramanian, S., (2017), Machine learning model for sequence-driven DNA G-quadruplex formation, *Sci. Rep.*, vol. 7, n°1, p. 14535.

Saranathan, N. et Vivekanandan, P., (2018), G-Quadruplexes: More Than Just a Kink in Microbial Genomes, *Trends Microbiol.*, vol. pii: S0966-842X, n°18, p. 30195-1.

Sauer, M. et Paeschke, K., (2017), G-quadruplex unwinding helicases and their function in vivo, *Biochem. Soc. Trans.*, vol. 45, n°5, p. 1173-1182.

Saxena, S., Miyoshi, D. et Sugimoto, N., (2010), Sole and Stable RNA Duplexes of G-Rich Sequences Located in the 5′-Untranslated Region of Protooncogenes, *Biochemistry*, vol. 49, n°33, p. 7190-7201.

Scaria, V., Hariharan, M., Arora, A. et Maiti, S., (2006), Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences, *Nucleic Acids Res.*, vol. 34, n°Web Server issue, p. W683-685.

Schaeffer, C., Bardoni, B., Mandel, J.-L., Ehresmann, B., Ehresmann, C. et Moine, H., (2001), The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif, *EMBO J.*, vol. 20, n°17, p. 4803-4813.

Schaffitzel, C., Berger, I., Postberg, J., Hanes, J., Lipps, H. J. et Plückthun, A., (2001), In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, n°15, p. 8572-8577.

Schludi, M. H. et Edbauer, D., (2017), Targeting RNA G-quadruplexes as new treatment strategy for C9orf72 ALS/FTD, *EMBO Mol. Med.*, vol. 10, n°1, p. 4-6.

Serikawa, T., Spanos, C., von Hacht, A., Budisa, N., Rappsilber, J. et Kurreck, J., (2017), Comprehensive identification of proteins binding to RNA G-quadruplex motifs in the 5' UTR of tumor-associated mRNAs, *Biochimie*, vol. 144, p. 169-184.

Shafer, R. H. et Smirnov, I., (2000), Biological aspects of DNA/RNA quadruplexes, *Biopolymers*, vol. 56, n°3, p. 209-227.

Shahid, R., Bugaut, A. et Balasubramanian, S., (2010), The BCL-2 5' untranslated region contains an RNA G-quadruplex-forming motif that modulates protein expression., *Biochemistry*, vol. 49, n°38, p. 8300-8306.

Shannon, R. D., (1976), Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, *Acta Cryst. A*, vol. 32, n°5, p. 751-767.

Sharp, A., Crabb, S. J., Cutress, R. I., Brimmell, M., Wang, X., Packham, G. et Townsend, P. A., (2004), BAG-1 in carcinogenesis, *Expert Rev. Mol. Med.*, vol. 6, n°7, p. 1-15.

Skeen, V. R., Collard, T. J., Southern, S. L., Greenhough, A., Hague, A., Townsend, P. A., … Williams, A. C., (2013), BAG-1 suppresses expression of the key regulatory cytokine transforming growth factor β (TGF-β1) in colorectal tumour cells, *Oncogene*, vol. 32, n°38, p. 4490-4499.

Soemedi, R., Cygan, K. J., Rhine, C. L., Glidden, D. T., Taggart, A. J., Lin, C.-L., … Fairbrother, W. G., (2017), The effects of structure on pre-mRNA processing and stability, *Methods*, vol. 125, p. 36-44.

Sofola, O. A., Jin, P., Qin, Y., Duan, R., Liu, H., de Haro, M., … Botas, J., (2007), RNA-Binding Proteins hnRNP A2/B1 and CUGBP1 Suppress Fragile X CGG Premutation Repeat-Induced Neurodegeneration in a Drosophila Model of FXTAS, *Neuron*, vol. 55, n°4, p. 565-571.

Sonenberg, N. et Hinnebusch, A. G., (2009), Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets, *Cell*, vol. 136, n°4, p. 731-745.

Song, J., Perreault, J.-P., Topisirovic, I. et Richard, S., (2016), RNA G-quadruplexes and their potential regulatory roles in translation, *Translation*, vol. 4, n°2, p. e1244031.

Song, J., Takeda, M. et Morimoto, R. I., (2001), Bag1-Hsp70 mediates a physiological stress signalling pathway that regulates Raf-1/ERK and cell growth, *Nat. Cell Biol.*, vol. 3, n°3, p. 276-282.

Soukup, G. A. et Breaker, R. R., (1999), Relationship between internucleotide linkage geometry and the stability of RNA., *RNA*, vol. 5, n°10, p. 1308-1325.

Southern, S. L., Collard, T. J., Urban, B. C., Skeen, V. R., Smartt, H. J., Hague, A., … Williams, A. C., (2012), BAG-1 interacts with the p50-p50 homodimeric NF-κB complex: implications for colorectal carcinogenesis, *Oncogene*, vol. 31, n°22, p. 2761-2772.

Sriram, A., Bohlen, J. et Teleman, A. A., (2018), Translation acrobatics : how cancer cells exploit alternate modes of translational initiation, *EMBO Rep.*, vol. 19, n°10, p. e45947.

Stefanovic, S., Bassell, G. J. et Mihailescu, M. R., (2015), G quadruplex RNA structures in PSD-95 mRNA : potential regulators of miR-125a seed binding site accessibility, *RNA*, vol. 21, n°1, p. 48-60.

Stefl, R., Oberstrass, F. C., Hood, J. L., Jourdan, M., Zimmermann, M., Skrisovska, L., … Allain, F. H.-T., (2010), The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove, *Cell*, vol. 143, n°2, p. 225-237.

Stegle, O., Payet, L., Mergny, J.-L., MacKay, D. J. C. et Huppert, J. L., (2009), Predicting and understanding the stability of G-quadruplexes, *Bioinformatics*, vol. 25, n°12, p. i374-i1382.

Subramanian, M., Rage, F., Tabet, R., Flatter, E., Mandel, J.-L. et Moine, H., (2011), G-quadruplex RNA structure as a signal for neurite mRNA targeting., *EMBO Rep.*, vol. 12, n°7, p. 697-704.

Sun, D. et Hurley, L. H., (2010), Biochemical Techniques for the Characterization of G-Quadruplex Structures: EMSA, DMS Footprinting, and DNA Polymerase Stop Assay, *Methods Mol. Biol.*, vol. 608, p. 65-79.

Suzuki, M. et Mizuno, A., (2004), A novel human Cl(-) channel family related to Drosophila flightless locus, *J. Biol. Chem.*, vol. 279, n°21, p. 22461-22468.

Swiatkowska, A., Kosman, J. et Juskowiak, B., (2016), FRET study of G-quadruplex forming fluorescent oligonucleotide probes at the lipid monolayer interface, *Spectrochim. Acta, Part A*, vol. 152, p. 614-621.

Takahama, K. et Oyoshi, T., (2013), Specific binding of modified RGG domain in TLS/FUS to G-quadruplex RNA : tyrosines in RGG domain recognize 2'-OH of the riboses of loops in G-quadruplex., *J. Am. Chem. Soc.*, vol. 135, n°48, p. 18016-18019.

Takahashi, S., Chelobanov, B., Kim, K. T., Kim, B. H., Stetsenko, D. et Sugimoto, N., (2018), Design and Properties of Ligand-Conjugated Guanine Oligonucleotides for Recovery of Mutated G-Quadruplexes, *Molecules*, vol. 23, n°12, p. E3228.

Takayama, S., Sato, T., Krajewski, S., Kochel, K., Irie, S., Millan, J. A. et Reed, J. C., (1995), Cloning and functional analysis of BAG-1: a novel Bcl-2-binding protein with anti-cell death activity, *Cell*, vol. 80, n°2, p. 279-284.

Taliaferro, J. M., Lambert, N. J., Sudmant, P. H., Dominguez, D., Merkin, J. J., Alexis, M. S., … Burge, C. B., (2016), RNA sequence context effects measured in vitro predict in vivo protein binding and regulation, *Mol Cell*, vol. 64, n°2, p. 294-306.

Tang, C.-F. et Shafer, R. H., (2006), Engineering the Quadruplex Fold, *J. Am. Chem. Soc.*, vol. 128, n°17, p. 5966-5973.

Tang, S.-C., (2002), BAG-1, an anti-apoptotic tumour marker, *IUBMB Life*, vol. 53, n°2, p. 99-105.

Tang, W., Robles, A. I., Beyer, R. P., Gray, L. T., Nguyen, G. H., Oshima, J., … Monnat, R. J., (2016), The Werner syndrome RECQ helicase targets G4 DNA in human cells to modulate transcription, *Hum. Mol. Genet.*, vol. 25, n°10, p. 2060-2069.

Taylor, S. C., Carbonneau, J., Shelton, D. N. et Boivin, G., (2015), Optimization of Droplet Digital PCR from RNA and DNA extracts with direct comparison to RT-qPCR: Clinical implications for quantification of Oseltamivir-resistant subpopulations, *J. Virol. Methods*, vol. 224, p. 58-66.

Terenin, I. M., Smirnova, V. V., Andreev, D. E., Dmitriev, S. E. et Shatsky, I. N., (2017), A researcher's guide to the galaxy of IRESs, *Cell. Mol. Life Sci.*, vol. 74, n°8, p. 1431-1455.

Thandapani, P., Song, J., Gandin, V., Cai, Y., Rouleau, S. G., Garant, J.-M., … Richard, S., (2015), Aven recognition of RNA G-quadruplexes regulates translation of the mixed lineage leukemia protooncogenes, *ELife*, vol. 4, p. e06234.

The Gene Ontology Consortium, (2017), Expansion of the Gene Ontology knowledgebase and resources, *Nucleic Acids Res.*, vol. 45, n°D1, p. D331-D338.

Thompson, S., (2012), So You Want to Know if Your Message Has an IRES?, *Wiley Interdiscip Rev RNA*, vol. 3, n°5, p. 697-705.

Tippana, R., Xiao, W. et Myong, S., (2014), G-quadruplex conformation and dynamics are determined by loop length and sequence, *Nucleic Acids Res.*, vol. 42, n°12, p. 8106-8114.

Todd, A. K., Johnston, M. et Neidle, S., (2005), Highly prevalent putative quadruplex sequence motifs in human DNA, *Nucleic Acids Res.*, vol. 33, n°9, p. 2901-2907.

Tomasko, M., Vorlícková, M. et Sagi, J., (2009), Substitution of adenine for guanine in the quadruplex-forming human telomere DNA sequence G(3)(T(2)AG(3))(3), *Biochimie*, vol. 91, n°2, p. 171-179.

Townsend, A. P., Stephanou, A., Packham, G. et Latchman, S. D., (2005), BAG-1 : a multi-functional pro-survival molecule., *Int. J. Biochem. Cell Biol.*, vol. 37, n°2, p. 251-259.

Townsend, P. A., Cutress, R. I., Sharp, A., Brimmell, M. et Packham, G., (2003), BAG-1 : a multifunctional regulator of cell growth and survival, *Biochim. Biophys. Acta*, vol. 1603, n°2, p. 83-98.

Trachman, R. J., Demeshkina, N. A., Lau, M. W. L., Panchapakesan, S. S. S., Jeng, S. C. Y., Unrau, P. J. et Ferré-D'Amaré, A. R., (2017), Structural basis for high-affinity fluorophore binding and activation by RNA Mango, *Nat. Chem. Biol.*, vol. 13, n°7, p. 807-813.

Varadaraj, K. et Skinner, D. M., (1994), Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases, *Gene*, vol. 140, n°1, p. 1-5.

Varizhuk, A., Ischenko, D., Smirnov, I., Tatarinova, O., Severov, V., Novikov, R., … Pozmogova, G., (2014), An Improved Search Algorithm to Find G-Quadruplexes in Genome Sequences, *BioRxiv*, p. 001990.

Varizhuk, A., Ischenko, D., Tsvetkov, V., Novikov, R., Kulemin, N., Kaluzhny, D., … Pozmogova, G., (2017), The expanding repertoire of G4 DNA structures, *Biochimie*, vol. 135, p. 54-62.

Vasilyev, N., Polonskaia, A., Darnell, J. C., Darnell, R. B., Patel, D. J. et Serganov, A., (2015), Crystal structure reveals specific recognition of a G-quadruplex RNA by a β-turn in the RGG motif of FMRP, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, n°39, p. E5391-5400.

Vattem, K. M. et Wek, R. C., (2004), Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, n°31, p. 11269-11274.

Verma, A., Halder, K., Halder, R., Yadav, V. K., Rawal, P., Thakur, R. K., … Chowdhury, S., (2008), Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species, *J. Med. Chem.*, vol. 51, n°18, p. 5641-5649.

Víglaský, V., Bauer, L. et Tlucková, K., (2010), Structural features of intra- and intermolecular G-quadruplexes derived from telomeric repeats, *Biochemistry*, vol. 49, n°10, p. 2110-2120.

von Hacht, A., Seifert, O., Menger, M., Schütze, T., Arora, A., Konthur, Z., … Kurreck, J., (2014), Identification and characterization of RNA guanine-quadruplex binding proteins., *Nucleic Acids Res.*, vol. 42, n°10, p. 6630-6644.

Vourekas, A., Zheng, K., Fu, Q., Maragkakis, M., Alexiou, P., Ma, J., … Wang, J. P., (2015), The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing., *Genes Dev.*, vol. 29, n°6, p. 617-629.

Vummidi, B. R., Alzeer, J. et Luedtke, N. W., (2013), Fluorescent Probes for G-Quadruplex Structures, *ChemBioChem*, vol. 14, n°5, p. 540-558.

Waldron, J. A., Raza, F. et Le Quesne, J., (2018), eIF4A alleviates the translational repression mediated by classical secondary structures more than by G-quadruplexes, *Nucleic Acids Res.*, vol. 46, n°6, p. 3075-3087.

Waller, Z. A. E., Sewitz, S. A., Hsu, S.-T. D. et Balasubramanian, S., (2009), A Small Molecule That Disrupts G-Quadruplex DNA Structure and Enhances Gene Expression, *J. Am. Chem. Soc.*, vol. 131, n°35, p. 12628-12633.

Wang, H. G., Takayama, S., Rapp, U. R. et Reed, J. C., (1996), Bcl-2 interacting protein, BAG-1, binds to and activates the kinase Raf-1, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 93, n°14, p. 7063-7068.

Wang, T. E., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., … Burge, B. C., (2008), Alternative isoform regulation in human tissue transcriptomes., *Nature*, vol. 456, n°7221, p. 470-476.

Wanrooij, H. P., Uhler, P. J., Simonsson, T., Falkenberg, M. et Gustafsson, M. C., (2010), G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation., *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, n°37, p. 16072-16077.

Watson, J., Baker, T., Bell, S., Gann, A., Levine, M. et Losick, R., (2009), *Biologie moléculaire du gène*, Pearson Education France, 688 p.

Watson, J. D. et Crick, F. H. C., (1953), Genetical Implications of the Structure of Deoxyribonucleic Acid, *Nature*, vol. 171, n°4361, p. 964.

Weinrich, T., Jaumann, E. A., Scheffer, U., Prisner, T. F. et Göbel, M. W., (2018), A Cytidine Phosphoramidite with Protected Nitroxide Spin Label: Synthesis of a Full-Length TAR RNA and Investigation by In-Line Probing and EPR Spectroscopy, *Chemistry – A European Journal*, vol. 24, n°23, p. 6202-6207.

Weldon, C., Behm-Ansmant, I., Hurley, L. H., Burley, G. A., Branlant, C., Eperon, I. C. et Dominguez, C., (2017a), Identification of G-quadruplexes in long functional RNAs using 7-deazaguanine RNA, *Nat. Chem. Biol.*, vol. 13, n°1, p. 18-20.

Weldon, C., Dacanay, J. G., Gokhale, V., Boddupally, P. V. L., Behm-Ansmant, I., Burley, G. A., … Eperon, I. C., (2017b), Specific G-quadruplex ligands modulate the alternative splicing of Bcl-X, *Nucleic Acids Res.*, vol. 46, n°2, p. 886-896.

Weldon, C., Eperon, I. C. et Dominguez, C., (2016), Do we know whether potential G-quadruplexes actually form in long functional RNA molecules?, *Biochem. Soc. Trans.*, vol. 44, n°6, p. 1761-1768.

Wells, S. E., Hughes, J. M., Igel, A. H. et Ares, M., (2000), Use of dimethyl sulfate to probe RNA structure in vivo, *Meth. Enzymol.*, vol. 318, p. 479-493.

Weng, H.-Y., Huang, H.-L., Zhao, P.-P., Zhou, H. et Qu, L.-H., (2012), Translational repression of cyclin D3 by a stable G-quadruplex in its 5’ UTR : implications for cell cycle regulation, *RNA Biol.*, vol. 9, n°8, p. 1099-1109.

Wiegering, A., Uthe, F. W., Jamieson, T., Ruoss, Y., Hüttenrauch, M., Küspert, M., … Eilers, M., (2015), Targeting translation initiation bypasses signaling crosstalk mechanisms that maintain high MYC levels in colorectal cancer, *Cancer Discov*, vol. 5, n°7, p. 768-781.

Wieland, M. et Hartig, J. S., (2007), RNA Quadruplex-Based Modulation of Gene Expression, *Chem. Biol.*, vol. 14, n°7, p. 757-763.

Wolfe, L. A., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, K. V., Sanghvi, R. V., … Wendel, H.-G., (2014), RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer., *Nature*, vol. 513, n°7516, p. 65-70.

Wong, H. M., Stegle, O., Rodgers, S. et Huppert, J. L., (2010), A toolbox for predicting g-quadruplex formation and stability, *J. Nucleic Acids*, vol. 2010, p. 564946.

Wood, J., Lee, S. S. et Hague, A., (2009), Bag-1 proteins in oral squamous cell carcinoma, *Oral Oncol.*, vol. 45, n°2, p. 94-102.

Xiao, C.-D., Ishizuka, T. et Xu, Y., (2017), Antiparallel RNA G-quadruplex Formed by Human Telomere RNA Containing 8-Bromoguanosine, *Sci. Rep.*, vol. 7, n°1, p. 6695.

Xiao, C.-D., Shibata, T., Yamamoto, Y. et Xu, Y., (2018), An intramolecular antiparallel G-quadruplex formed by human telomere RNA, *Chem. Commun.*, vol. 54, n°32, p. 3944-3946.

Xiao, S., Zhang, J.-Y., Zheng, K.-W., Hao, Y.-H. et Tan, Z., (2013), Bioinformatic analysis reveals an evolutional selection for DNA : RNA hybrid G-quadruplex structures as putative transcription regulatory elements in warm-blooded animals., *Nucleic Acids Res.*, vol. 41, n°22, p. 10379-10390.

Xiong, J., Chen, J., Chernenko, G., Beck, J., Liu, H., Pater, A. et Tang, S.-C., (2003), Antisense BAG-1 sensitizes HeLa cells to apoptosis by multiple pathways, *Biochem. Biophys. Res. Commun.*, vol. 312, n°3, p. 585-591.

Xu, S., Li, Q., Xiang, J., Yang, Q., Sun, H., Guan, A., … Tang, Y., (2015), Directly lighting up RNA G-quadruplexes from test tubes to living human cells, *Nucleic Acids Res.*, vol. 43, n°20, p. 9575-9586.

Xue, S., Tian, S., Fujii, K., Kladwang, W., Das, R. et Barna, M., (2015), RNA regulons in Hox 5′UTRs confer ribosome specificity to gene regulation, *Nature*, vol. 517, n°7532, p. 33-38.

Xue, Y., Liu, J., Zheng, K., Kan, Z., Hao, Y. et Tan, Z., (2011), Kinetic and thermodynamic control of G-quadruplex folding, *Angew. Chem., Int. Ed. Engl.*, vol. 50, n°35, p. 8046-8050.

Yadav, K. V., Abraham, K. J., Mani, P., Kulshrestha, R. et Chowdhury, S., (2008), QuadBase: genome-wide database of G4 DNA--occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes., *Nucleic Acids Res.*, vol. 36, p. D381-D385.

Yamaguchi, K., Asakura, K., Imamura, M., Kawai, G., Sakamoto, T., Furihata, T., … Higashi, K., (2018), Polyamines stimulate the CHSY1 synthesis through the unfolding of the RNA G-quadruplex at the 5'-untraslated region, *Biochem. J.*, vol. 475, n°23, p. 3797-3812.

Yaman, I., Fernandez, J., Liu, H., Caprara, M., Komar, A. A., Koromilas, A. E., … Hatzoglou, M., (2003), The Zipper Model of Translational Control: A Small Upstream ORF Is the Switch that Controls Structural Remodeling of an mRNA Leader, *Cell*, vol. 113, n°4, p. 519-531.

Yang, S. Y., Lejault, P., Chevrier, S., Boidot, R., Robertson, A. G., Wong, J. M. Y. et Monchaud, D., (2018), Transcriptome-wide identification of transient RNA G-quadruplexes in human cells, *Nat .Commun.*, vol. 9, n°1, p. 4730.

Yang, X., Chernenko, G., Hao, Y., Ding, Z., Pater, M. M., Pater, A. et Tang, S. C., (1998), Human BAG-1/RAP46 protein is generated as four isoforms by alternative translation initiation and overexpressed in cancer cells, *Oncogene*, vol. 17, n°8, p. 981-989.

Ying, L., Green, J. J., Li, H., Klenerman, D. et Balasubramanian, S., (2003), Studies on the structure and dynamics of the human telomeric G quadruplex by single-molecule fluorescence resonance energy transfer, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, n°25, p. 14629-14634.

Zaccaria, F. et Fonseca Guerra, C., (2018), RNA versus DNA G-Quadruplex: The Origin of Increased Stability, *Chemistry*, vol. 24, n°61, p. 16315-16322.

Zamiri, B., Reddy, K., Macgregor, R. B. et Pearson, C. E., (2014), TMPyP4 porphyrin distorts RNA G-quadruplex structures of the disease-associated r(GGGGCC)n repeat of the C9orf72 gene and blocks interaction of RNA-binding proteins, *J. Biol. Chem.*, vol. 289, n°8, p. 4653-4659.

Zeiner, M. et Gehring, U., (1995), A protein that interacts with members of the nuclear hormone receptor family: identification and cDNA cloning, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 92, n°25, p. 11465-11469.

Zeng, L., Qian, J., Luo, X., Zhou, A., Zhang, Z. et Fang, Q., (2018), CHSY1 promoted proliferation and suppressed apoptosis in colorectal cancer through regulation of the NFκB and/or caspase-3/7 signaling pathway, *Oncol Lett*, vol. 16, n°5, p. 6140-6146.

Zeraati, M., Moye, A. L., Wong, J. W. H., Perera, D., Cowley, M. J., Christ, D. U., … Dinger, M. E., (2017), Cancer-associated noncoding mutations affect RNA G-quadruplex-mediated regulation of gene expression, *Sci. Rep.*, vol. 7, n°1, p. 708.

Zhang, A. Y. Q. et Balasubramanian, S., (2012), The Kinetics and Folding Pathways of Intramolecular G-Quadruplex Nucleic Acids, *J. Am. Chem. Soc.*, vol. 134, n°46, p. 19297-19308.

Zhang, A. Y. Q., Bugaut, A. et Balasubramanian, S., (2011a), A sequence-independent analysis of the loop length dependence of intramolecular RNA G-quadruplex stability and topology, *Biochemistry*, vol. 50, n°33, p. 7251-7258.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., … Cptac, the N., (2014), Proteogenomic characterization of human colon and rectal cancer, *Nature*, vol. 513, n°7518, p. 382-387.

Zhang, D.-H., Fujimoto, T., Saxena, S., Yu, H.-Q., Miyoshi, D. et Sugimoto, N., (2010a), Monomorphic RNA G-Quadruplex and Polymorphic DNA G-Quadruplex Structures Responding to Cellular Environmental Factors, *Biochemistry*, vol. 49, n°21, p. 4554-4563.

Zhang, P., He, D., Xu, Y., Hou, J., Pan, B.-F., Wang, Y., … Chen, Y., (2017), Genome-wide identification and differential analysis of translational initiation, *Nat. Commun.*, vol. 8, n°1, p. 1749.

Zhang, W. et Chen, S.-J., (2002), RNA hairpin-folding kinetics, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, n°4, p. 1931-1936.

Zhang, Y., Gaetano, M. C., Williams, R. K., Bassell, J. G. et Mihailescu, R. M., (2011b), FMRP interacts with G-quadruplex structures in the 3'-UTR of its dendritic target Shank1 mRNA., *RNA Biol.*, vol. 11, n°11, p. 1364-1374.

Zhang, Z., Dai, J., Veliath, E., Jones, R. A. et Yang, D., (2010b), Structure of a two-G-tetrad intramolecular G-quadruplex formed by a variant human telomeric sequence in K+ solution: insights into the interconversion of human telomeric G-quadruplex structures, *Nucleic Acids Res.*, vol. 38, n°3, p. 1009-1021.

Zheng, K., Xiao, S., Liu, J., Zhang, J., Hao, Y. et Tan, Z., (2013), Co-transcriptional formation of DNA : RNA hybrid G-quadruplex and potential function as constitutional cis element for transcription control., *Nucleic Acids Res.*, vol. 41, n°10, p. 5533-5541.

Zhou, B., Liu, C., Geng, Y. et Zhu, G., (2015), Topology of a G-quadruplex DNA formed by *C9orf72* hexanucleotide repeats associated with ALS and FTD, *Sci. Rep.*, vol. 5, p. 16673.

Zuker, M., Mathews, D. H. et Turner, D. H., (1999), Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide, In : J. BARCISZEWSKI & B. F. C. CLARK (éd.), *RNA Biochemistry and Biotechnology*, Springer Netherlands, Dordrecht, p. 11-43.

# REMERCIEMENTS

Je tiens en premier lieu à remercier les membres de mon jury qui ont accepté d'évaluer ma thèse. Votre expertise scientifique ainsi que vos critiques et commentaires constructifs sont grandement appréciés.

Je remercie énormément mon directeur de recherche Jean-Pierre Perreault de m'avoir accueilli dans son laboratoire dès 2009 alors que j'étais une étudiante de 1re année du bacc. à la recherche d'un stage. Dès cette première rencontre, je me suis sentie à ma place. Merci d'avoir eu confiance en moi et de m'avoir donné le support, l'environnement, l'autonomie et la flexibilité pour apprendre à faire la recherche par moi-même. Jean-Pierre m'a laissé une grande liberté, sachant quand me pousser à me dépasser et en me donnant uniquement les corrections et les conseils nécessaires aux moments opportuns. Je tiens également à remercier mon co-directeur Martin Bisaillon. Tout les deux, avec votre leadership par l'exemple, par vos méthodes de travail, et le partage de vos anecdotes et de votre expérience avez été d'excellents modèles.

Merci aux membres de mon comité d'encadrement : Nathalie Rivard et Michelle Scott pour l'accompagnement et leur regard extérieur durant tout mon parcours aux études doctorales. Je remercie également les membres de mon jury d'examen général Éric Massé, Robert Day, et Stefania Millevoi qui ont contribué par leurs évaluations au cheminement de mon projet.

Merci à tous mes collègues du laboratoire depuis les premiers jours jusqu'aux derniers, dont le travail portait sur les G4, les viroïdes ou les ribozymes. J'ai apprécié être avec eux tous les jours, ainsi que toutes les discussions scientifiques ou non que nous avons partagé à l'intérieur comme à l'extérieur du labo. Un merci plus particulier à Jean-Denis Beaudoin qui a été mon exemple no.1 d'étudiant gradué et qui m'a transmis la passion des G4 et de l'*in-line*. Il y a également une place particulière dans mon cœur à la gang de ma « génération » labo Perreault : Samuel Rouleau, Tamara Giguère, Jean-Michel Garant et Lubos Bauer, pour tout le temps passé ensemble à la paillasse, au 5 à 7 ou en voyage ! Clin d'œil aussi à mes amis du labo « Perr-aillon ».

Merci aux professionnels de recherche passés, Dominique Lévesque et Julie Motard pour m'avoir enseigné les rudiments des manips. Je remercie tout particulièrement François Bolduc, sans lui rien ne fonctionnerait dans le labo et j'aurais accumulé encore plus de boîtes dans le congélo… Je remercie également les stagiaires qui ont contribué à mes projets : Mahdi Abdelhamid Laaref, Josiann Normandeau-Guimond, Cameron Levins et Hélène Cossette-Roberge.

Je suis reconnaissante envers la communauté scientifique du Ribo-club et des G-quadruplexes Meetings pour les opportunités de présenter mes résultats, de rencontrer de grands chercheurs et la participation à la vie scientifique en congrès locaux ou internationaux. Cela a agrandi mon esprit scientifique.

Finalement, je tiens à remercier toute ma famille : mes parents Maryse et Pierre-Guy, ma sœur Zoé, ainsi que mon amoureux Bruno-Pier et sa famille. J'ajoute à ce groupe tous mes amis et ma « famille » d'ultimate. Merci pour votre support inconditionnel, votre soutien physique et émotionnel qui m'a permis de garder l'équilibre tout au long. Je vous aime.

# ANNEXES

**ANNEXE 1 Tableau A1 Outils de prédictions des G4 classés par catégorie**

| Nom | Type de prédiction | Chevauchement | Caractéristiques non canoniques | | | Score | Outil Web | Réf. |
|---|---|---|---|---|---|---|---|---|
| | | | Renfl. | Mésapp. | Boucles (>12 nt) | | | |
| **Recherche motif canonique** | | | | | | | | |
| Quadparser | Motif canonique | Oui | Non | Non | Non | Non | URL n'existe plus *http://www.quadruplex.org* | Huppert 2005, Todd 2005 |
| QGRS mapper | Motif canonique modifiable | Oui | Non | Non | Non | G-score | *http://bioinformatics.ramapo.edu/QGRS/index.php* | Kikin, 2006 |
| QGRS-H | Conservation du motif | Oui | Non | Non | Non | G-score | *http://quadruplex.ramapo.edu/qgrs/app/start* | Menendez, 2012 |
| Quadfinder | Motif canonique modifiable | Oui | Non | Non | Non | Non | URL n'existe plus *http://miracle.igib.res.in/quadfinder/* | Scaria, 2006 |
| Quadbase | Motif canonique modifiable | Oui | Non | Non | Non | Non | n'est plus disponible depuis la v.2 | Yadav, 2008 |
| nBMST (*non-B DNA Motif search too*l) | Motif canonique | Oui | Non | Non | Non | Non | *http://nonb.abcc.ncifcrf.gov/apps/nBMST/default/* | Cer, 2013 |
| G4IPDB (G4-predictor v.1-2) | Motif canonique | Oui | Non | Non | Oui, | cG/cC | *http://bsbe.iiti.ac.in/bsbe/ipdb/index.php* | Mishra, 2016 |

| Nom | Type de prédiction | Chevauchement | Caractéristiques non canoniques | | | Score | Outil Web | Réf. |
|---|---|---|---|---|---|---|---|---|
| | | | Renfl. | Mésapp. | Boucles (>12 nt) | | | |
| **Stabilité structure secondaire** | | | | | | | | |
| Quadpredict | *Gaussian process regression* | Non | Non | Non | Non | Non | URL n'existe plus *http://www.quadruplex.org* | Stegle, 2009 |
| RNAfold G4 | Stabilité structure secondaire + recherche motif | Non | Non | Non | Non | Non | RNAfold webserver, ajouter G-quadruplex prediction dans les options avancées *http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi* | Lorenz, 2013, 2D meet 4D) |
| **Recherche motifs irréguliers** | | | | | | | | |
| Quadbase2/TetraplexFinder | Expression régulière (Regex) | Oui | Oui | Non | Oui | Non | *http://quadbase.igib.res.in/* | Dhapola, 2016 |
| imGQfinder | Expression régulière (Regex) | Oui | Oui* * 1 seule caract. canonique à la fois | | non - | Non | *http://imgqfinder.niifhm.ru* | Varizhuk, 2017 |
| pqsfinder | Expression régulière (Regex) | Oui | Oui | Oui | Non | Oui | Non, R/Bioconductor package disponible à : *http://bioconductor.org/packages/pqsfinder/* | Hon, 2017 |
| G4Catchall (G4C) | Expression régulière (Regex) | Oui | Oui 2 max. | Oui 2 max | Oui 1 max | Non | *http://github.com/odoluca/G4Catchall* | Doluca, 2018 |
| **Contexte et densité** | | | | | | | | |
| G4P-calculator | Densité fenêtres défilantes | Oui | Oui | Oui | Oui | Non | Programme téléchargeable : *http://depts.washington.edu/maizels9/G4calc.php* | Eddy and Maizels 2006, 2008 |
| cG/cC score | Densité G et C dans une fenêtre | N/A | Oui | Oui | Oui | cG/cC | Non, pas dans la publication originale | Beaudoin, 2014 |
| G4Hunter | Fenêtre défilante | Non | Oui | Oui | Oui | G4H score | Non, script R disponible | Bedrat, 2016 |

| Nom | Type de prédiction | Chevauchement | Caractéristiques non canoniques | | | Score | Outil Web | Réf. |
|---|---|---|---|---|---|---|---|---|
| | | | Renfl. | Mésapp. | Boucles (>12 nt) | | | |
| **Apprentissage automatisé** | | | | | | | | |
| G4-HMM | Motifs canoniques discriminés avec *Hidden Markov models* | Non | Non | Non | Non | Z-score | Non, programmes C++ et Perl disponibles à : *http://tcs.cira.kyoto-u.ac.jp/~ykato/program/g4hmm/* | Yano & Kato, 2014 |
| G4 predictor project | *Support Vector Machine with String Kernel model* | Oui | Oui | Oui | Oui | Non | *http://g4predictor.appspot.com/* | Tradigo, 2014 |
| G4screener | Réseau de neurones | Oui | Oui | Oui | Oui | G4NN G4H, cGcC | *http://scottgroup.med.usherbrooke.ca/G4RNA_screener/* | Garant, 2017 |
| Quadron | *Tree-based gradient boosting machines (GBMs)* | Oui | Non | Non | Non | Quadron score | Non, code source et programme disponible à : *http://quadron.atgcdynamics.org* | Sahakyan, 2017 |

**ANNEXE 2 Supplementary data Article 2**

# Supplementary data

## Article 2 – New scoring system to identify RNA G-quadruplex folding

**Supplementary Tables S1-S6**

**Table S1** Oligodeoxynucleotides used to synthesize PG4 candidates

**Table S2** RNA sequences of the PG4 probed *in vitro* by in-line probing

**Table S3** Full length 3'-UTR RNA sequences of candidates

**Table S4** Oligodeoxynucleotides used to build full length 3'-UTR and directed mutagenesis

**Table S5** Area under the curve of the different predictive parameters

**Table S6** Table of sensitivity and specificity percentages for the different cG/cC score thresholds.

**Supplementary Figures and Legends S1-S27**

**Figures S1-S26** In-line probing results of each candidate.

**Figure S27** Secondary structures of MAP3K11 PG4 WT short and long candidates.

**Supplementary table S1** Oligodeoxynucleotides used to synthesize PG4 candidates

| ***in vitro* PG4 candidates** | | | **Oligo 1 foward (5'-3')** | **Oligo 2 reverse (5'-3')** |
|---|---|---|---|---|
| NCAM2 | Short | WT | TAATACGACTCACTATAGGG | CCGCCGCCGCCGCTCCCGCGGTGCCCCAGCCGCCCGCAGC CCGCGCGCTCCTCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | CCGCCGCCGCCGCTCTCGCGGTGCTTCAGCCGCTCGCAGC TCGCGCGCTCCTCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGATAGTGCGGCAAGAGCGG AGCTTGCAGTCACTTTGCGAGGAGGAGCGCGC | GAGAACCTTTCGCTGCCCCGGCCGCCTCTGCTAGAGCCGC CGCCGCCGCTCCCGCGGTGCCCCAGCCGCCCGCAGCCCGC GCGCTCCTCCTCGCAAAGTG |
| | | G/A-mutant | TAATACGACTCACTATAGGGATAGTGCGGCAAGAGCGG AGCTTGCAGTCACTTTGCGAGGAGGAGCGCGC | GAGAACCTTTCGCTGCCCCGGCCGCCTCTGCTAGAGCCGC CGCCGCCGCTCTCGCGGTGCTTCAGCCGCTCGCAGCTCGC GCGCTCCTCCTCGCAAAGTG |
| BARHL1 | Short | WT | TAATACGACTCACTATAGGG | CCCCAAAAGCTCCCCACCCACCCCCAACCCCAGCCGCCGC TGCCCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | CCCCAAAAGCTCTTCACCCACTCTCAACTTCAGCCGCCGC TGCCCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGCCCGCCTTCCCCATGCCA GCCCGCAGCTAGGGGCAGGGGCAGCGGCGGCTG | GAAGTCATGGTGGCTAGCAGCCAAGCTGCGACCTGTCCTC CCCAAAAGCTCCCCACCCACCCCCAACCCCAGCCGCCGCT GCCCCTGCCC |
| | | G/A-mutant | TAATACGACTCACTATAGGGCCCGCCTTCCCCATGCCA GCCCGCAGCTAGGGGCAGGGGCAGCGGCGGCTG | GAAGTCATGGTGGCTAGCAGCCAAGCTGCGACCTGTCCTC CCCAAAAGCTCTTCACCCACTCTCAACTTCAGCCGCCGCT GCCCCTGCCC |
| FZD2 | Short | WT | TAATACGACTCACTATAGGG | CCCGGCTCCTTGGCGCCCCCCCCGCCCCCGCCCCCAACCC GGAGACTGCGCTTCTTCCCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | CCCGGCTCCTTGGCGCTTCCTTCGCTCTCGCTTCCAACCC GGAGACTGCGCTTCTTCCCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGAGGCGGCAGCCGCAGCGA GGAGGCGGCGGGGAAGAAGCGCAGTCTCC | AGTCATGGTGGCTAGCGCTGGCCGCCGCCCCCCACCCGGC TCCTTGGCGCCCCCCCCGCCCCCGCCCCCAACCCGGAGAC TGCGCTTCTTCCC |

| | | | | |
|---|---|---|---|---|
| | | G/A-mutant | TAATACGACTCACTATAGGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGTCTCC | AGTCATGGTGGCTAGCGCTGGCCGCCGCCCCCCACCCGGCTCCTTGGCGCTTCCTTCGCTCTCGCTTCCAACCCGGAGACTGCGCTTCTTCCCC |
| EBAG9 | Short | WT | TAATACGACTCACTATAGGG | TCAAAACCTGCCCCTCCCCCTCCCCGCCCGCCCGGCGGAGGCTCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | TCAAAACCTGCCCCTCCTTCTCCTTGCTTGCCCGGCGGAGGCTCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGAGCGCGCCTTGTGTGCGCGCGCGGCCCGCGGCAGCTCGGAGCCTCCGCC | GATCATAAACTTTCGAAGTCATGGTGGCTAGCGGTGGGAATCAAAACCTGCCCCTCCCCCTCCCCGCCCGCCCGGCGGAGGCTCCGAGCTGCCGCGG |
| | | G/A-mutant | TAATACGACTCACTATAGGGAGCGCGCCTTGTGTGCGCGCGCGGCCCGCGGCAGCTCGGAGCCTCCGCC | GATCATAAACTTTCGAAGTCATGGTGGCTAGCGGTGGGAATCAAAACCTGCCCCTCCTTCTCCTTGCTTGCCCGGCGGAGGCTCCGAGCTGCCGCGG |
| FXR1 | Short | WT | TAATACGACTCACTATAGGG | TATTACGTTAGGGATCCCACCCCACCCACCACCCCCCATCTCTGCCATATTTTGCCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | TATTACGTTAGGGATCtCACCtCACtCACCACtCtCCATCTCTGCCATATTTTGCCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGTTGCTGGCTATAGGAAATGTTATTTTGTTTTCAAAATATGGCAGAGATG | AATGTAACAAAAGCAGCTAATGCTTTCATAAAGAATATTACGTTAGGGATCCCACCCCACCCACCACCCCCCATCTCTGCCATATTTTG |
| | | G/A-mutant | TAATACGACTCACTATAGGGTTGCTGGCTATAGGAAATGTTATTTTGTTTTCAAAATATGGCAGAGATG | AATGTAACAAAAGCAGCTAATGCTTTCATAAAGAATATTACGTTAGGGATCTCACCTCACTCACCACTCTCCATCTCTGCCATATTTTG |
| LRP5 | Short | WT | TAATACGACTCACTATAGGG | CTGTACAAAGTTCTCCCAGCCCTGCCCACCCCATCACAGTTCACATTTCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | CTGTACAAAGTTCTCTCAGCTCTGCTCACTCCATCACAGTTCACATTTCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGAAAATAAATATAATTGGGATTTTAAAAACATGAGAAATGTGAACTGTGATGG | CTGTTTTACAAAATTAAGTTTATAAATATTTCTCCACTGTACAAAGTTCTCCCAGCCCTGCCCACCCCATCACAGTTCACATTTCTCATG |
| | | G/A-mutant | TAATACGACTCACTATAGGGAAAATAAATATAATTGGGATTTTAAAAACATGAGAAATGTGAACTGTGATGG | CTGTTTTACAAAATTAAGTTTATAAATATTTCTCCACTGTACAAAGTTCTCTCAGCTCTGCTCACTCCATCACAGTTCACATTTCTCATG |
| AASDHPPT | Short | WT | TAATACGACTCACTATAGGG | GCAGACTGGCCCACCGACAGCCCTCCCAGCCCGCCCCCTATAGTGAGTCGTATTA |

| | | | | |
|---|---|---|---|---|
| | | G/A-mutant | TAATACGACTCACTATAGGG | GCAGACTGGCCCACCGACAGCTCTCTCAGCTCGCTCTCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGATTACAGCTCTTAAGGCTAGAGTACTTAATACGACTCACTATAGGCTAGC | GTCCGTGCGCAGGGGACGGGCCGTCGCTACGCAGACTGGCCCACCGACAGCCCTCCCAGCCCGCCCCCGCTAGCCTATAGTGAGTCGTATTAAG |
| | | G/A-mutant | TAATACGACTCACTATAGGGATTACAGCTCTTAAGGCTAGAGTACTTAATACGACTCACTATAGGCTAGC | GTCCGTGCGCAGGGGACGGGCCGTCGCTACGCAGACTGGCCCACCGACAGCTCTCTCAGCTCGCTCTCGCTAGCCTATAGTGAGTCGTATTAAG |
| THRA1 | Short | WT | TAATACGACTCACTATAGGG | CCCACAGGCCACCCACCCCCTGCCACCCAGGCCCTAGGGCACAGCACCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | CCCACAGGCCAATTATTTTATGCCACCCAGGCCCTAGGGCACAGCACCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGTGGTGTGAAAGGCCAAGTGCTGAGGCGGGTATCATGGGTGCTGTGCCCTA | CCAAGAGACTGGGGTGGGCACACTGGCCCCCCCGGCACACCCACAGGCCACCCACCCCCTGCCACCCAGGCCCTAGGGCACAGCACCCATGATACC |
| | | G/A-mutant | TAATACGACTCACTATAGGGTGGTGTGAAAGGCCAAGTGCTGAGGCGGGTATCATGGGTGCTGTGCCCTA | CCAAGAGACTGGGGTGGGCACACTGGCCCCCCCGGCACACCCACAGGCCAATTATTTTATGCCACCCAGGCCCTAGGGCACAGCACCCATGATACC |
| DOC2B | Short | WT | TAATACGACTCACTATAGGG | CCCCGGGCGCGGCCCGGCCCGGCGCGACCCCGGCCCGGGGGCGGCTCAGCAGGCCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | CCCCGGGCGCGGCTCGGCTCGGCGCGACTTCGGCTCGGGGGCGGCTCAGCAGGCCCTATAGTGAGTCGTATTA |
| | | C/A-mutant | TAATACGACTCACTATAGGG | CCCCGTGCGCGTCCCGGCCCTTCTCGACCCCGTCCCGTGGGCTTCTCATCATTCCCTATAGTGAGTCGTATTA |
| | | GC/AA-mutant | TAATACGACTCACTATAGGG | CCCCGTGCGCGTCTCGGCTCTTCTCGACTTCGTCTCGTGGGCTTCTCATCATTCCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGCGATGCCCGCAGCCCCCGCCGCGCCCCGCCGGGCCTGCTGAGCCGCCCCCG | TCGAAGTCATGGTGGCTAGCGCAGGCAGCGCCGCCCCGCCCCGGGCGCGGCCCGGCCCGGCGCGACCCCGGCCCGGGGGCGGCTCAGCAGGCCC |
| | | G/A-mutant | TAATACGACTCACTATAGGGCGATGCCCGCAGCCCCCGCCGCGCCCCGCCGGGCCTGCTGAGCCGCCCCCG | TCGAAGTCATGGTGGCTAGCGCAGGCAGCGCCGCCCCGCCCCGGGCGCGGCTCGGCTCGGCGCGACTTCGGCTCGGGGGCGGCTCAGCAGGCCC |
| | | C/A-mutant | TAATACGACTCACTATAGGGCGATGCCCGCAGCCCCCGCCGCGCCCCGCCGGGAATGTTGAGAAGCCCACG | TCGAAGTCATGGTGGCTAGCGCAGGCAGCGCCGCCCCGCCCCGTGCGCGTCCCGGCCCTTCTCGACCCCGTCCCGTGGGCTTCTCAACATTCCC |

| | | | | |
|---|---|---|---|---|
| | | GC/AA-mutant | TAATACGACTCACTATAGGGCGATGCCCGCAGCCCCCG CCGCGCCCCGCCGGGAATGTTGAGAAGCCCACG | TCGAAGTCATGGTGGCTAGCGCAGGCAGCGCCGCCCCGCC CCGTGCGCGTCTCGGCTCTTCTCGACTTCGTCTCGTGGGC TTCTCAACATTCCC |
| TNFSF12 | Short | WT | TAATACGACTCACTATAGGG | CCTGCCTCACCGCCCCCCCATCCCGGGACCCGAGGGATCG GGGGAGGGGGAGCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | CCTGCCTCACCGCTTCCTCATCTCGGGACTCGAGGGATCG GGGGAGGGGGAGCCTATAGTGAGTCGTATTA |
| | | C/A-mutant | TAATACGACTCACTATAGGG | CCTGCCTCACCtCCCCCCCATCCCGGGACCCtAtttATCG tGtGAGtGtGAtCCTATAGTGAGTCGTATTA |
| | | GC/AA-mutant | TAATACGACTCACTATAGGG | CCTGCCTCACCTCTTCCTCATCTCGGGACTCTATTTATCG TGTGAGTGTGATCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGCCTCTCCCCGGCCCGATC CGCCCGCCGGCTCCCCCTCCCCCGATCCCTCG | TTCGAAGTCATGGTGGCTAGCGGGGGCGGGGGGCTGTGCC TGCCTCACCGCCCCCCCATCCCGGGACCCGAGGGATCGGG GGAGGGGGAGCC |
| | | G/A-mutant | TAATACGACTCACTATAGGGCCTCTCCCCGGCCCGATC CGCCCGCCGGCTCCCCCTCCCCCGATCCCTCG | TTCGAAGTCATGGTGGCTAGCGGGGGCGGGGGGCTGTGCC TGCCTCACCGCTTCCTCATCTCGGGACTCGAGGGATCGGG GGAGGGGGAGCC |
| | | C/A-mutant | TAATACGACTCACTATAGGGCCTCTCCCCGGCCCGATC CGCCCGCCGGATCACACTCACACGATAAATAG | TTCGAAGTCATGGTGGCTAGCGGGGGCGGGGGGCTGTGCC TGCCTCACCTCCCCCCCATCCCGGGACCCTATTTATCGTG TGAGTGTGATCC |
| | | GC/AA-mutant | TAATACGACTCACTATAGGGCCTCTCCCCGGCCCGATC CGCCCGCCGGATCACACTCACACGATAAATAG | TTCGAAGTCATGGTGGCTAGCGGGGGCGGGGGGCTGTGCC TGCCTCACCTCTTCCTCATCTCGGGACTCTATTTATCGTG TGAGTGTGATCC |
| MAP3K11 | Short | WT | TAATACGACTCACTATAGGG | CCTGGGCATCCGGGCCCTGGCCCTCAGCCCCAGACCCACG CCTCTCTGGGGAGCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | CCTGGGCATCCGGGCTCTGGCTCTCAGCTCCAGACTCACG CCTCTCTGGGGAGCCTATAGTGAGTCGTATTA |
| | | C/A-mutant | TAATACGACTCACTATAGGG | CCTTGGCATCCGGTCCCTGTCCCTCATCCCCAGACCCACG CCTCTCTGGTGATCCTATAGTGAGTCGTATTA |
| | | GC/AA-mutant | TAATACGACTCACTATAGGG | CCTTGGCATCCGGTCTCTGTCTCTCATCTCCAGACTCACG CCTCTCTGGTGATCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGCGAGATGCGGGGGGCCGG GAGACAACACTCCTGGCTCCCCAGAGAGGCGTG | CCACCCCCGCTGGCTGCCAAGGCCCTAGTCCCGGAACCTG GGCATCCGGGCCCTGGCCCTCAGCCCCAGACCCACGCCTC TCTGGGGAGCCAGG |

| | | | | |
|---|---|---|---|---|
| | | G/A-mutant | TAATACGACTCACTATAGGGCGAGATGCGGGGGGCCGGGAGACAACACTCCTGGCTCCCCAGAGAGGCGTG | CCACCCCCGCTGGCTGCCAAGGCCCTAGTCCCGGAACCTGGGCATCCGGGCTCTGGCTCTCAGCTCCAGACTCACGCCTCTCTGGGGAGCCAGG |
| | | C/A-mutant | TAATACGACTCACTATAGGGCGAGATGCGGGGGGCCGGGAGACAACACTCCTGGATCACCAGAGAGGCGTG | CCACCCCCGCTGGCTGCCAAGGCCCTAGTCCCGGAACCTTGGCATCCGGTCCCTGTCCCTCAACCCCAGACCCACGCCTCTCTGGTGATCCAGG |
| | | GC/AA-mutant | TAATACGACTCACTATAGGGCGAGATGCGGGGGGCCGGGAGACAACACTCCTGGATCACCAGAGAGGCGTG | CCACCCCCGCTGGCTGCCAAGGCCCTAGTCCCGGAACCTTGGCATCCGGTCTCTGTCTCTCAACTCCAGACTCACGCCTCTCTGGTGATCCAGG |
| TTYH1 | Short | WT | TAATACGACTCACTATAGGG | GCGAGGCTGTCTGCCCCCTCCCCTGCCAGCCCTACTCCCTAGTCTGCCCCCTCAGCTACTCCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | GCGAGGCTGTCTGCTCTCTCCTCTGCCAGCTCTACTCTCTAGTCTGCTCCCTCAGCTACTCCCTATAGTGAGTCGTATTA |
| | Long | WT | TAATACGACTCACTATAGGGTGCTCCCATTTCTGTCCTTGGCCTTGGGAGTAGCTGAGGG | AAGGATGGGACCGGCAGCCAGGGATGAAGGGTGCGAGGCGAGGCTGTCTGCCCCCTCCCCTGCCAGCCCTACTCCCTAGTCTGCCCCCTCAGCTACTCCC |
| | | G/A-mutant | TAATACGACTCACTATAGGGTGCTCCCATTTCTGTCCTTGGCCTTGGGAGTAGCTGAGGG | AAGGATGGGACCGGCAGCCAGGGATGAAGGGTGCGAGGCGAGGCTGTCTGCTCTCTCCTCTGCCAGCTCTACTCTCTAGTCTGCTCCCTCAGCTACTCCC |
| | | C/A-mutant-1-mutant | TAATACGACTCACTATAGGGTGCTCCCATTTCTGTCCTTGGCCTTGGGAGTAGCTGAGGG | AAGGATGTGACCGGCAGCCAGTGATGAAGTGTGCGATGCGATGCTGTCTGCCCCCTCCCCTGCCAGCCCTACTCCCTAGTCTGCCCCCTCAGCTACTCCC |
| | | GC/AA-mutant-1 | TAATACGACTCACTATAGGGTGCTCCCATTTCTGTCCTTGGCCTTGGGAGTAGCTGAGGG | AAGGATGTGACCGGCAGCCAGTGATGAAGTGTGCGATGCGATGCTGTCTGCTCTCTCCTCTGCCAGCTCTACTCTCTAGTCTGCTCCCTCAGCTACTCCC |
| | | C/A-mutant-2-mutant | TAATACGACTCACTATAGGGTGATCACATTTCTGTACTTGGAATTGGGAGTAGCTGAGGG | AAGGATGTGACCGGCAGCCAGTGATGAAGTGTGCGATGCGATGCTGTCTGCCCCCTCCCCTGCCAGCCCTACTCCCTAGTCTGCCCCCTCAGCTACTCCC |
| | | GC/AA-mutant-2 | TAATACGACTCACTATAGGGTGATCACATTTCTGTACTTGGAATTGGGAGTAGCTGAGGG | AAGGATGTGACCGGCAGCCAGTGATGAAGTGTGCGATGCGATGCTGTCTGCTCTCTCCTCTGCCAGCTCTACTCTCTAGTCTGCTCCCTCAGCTACTCCC |
| TTYH1-LRP5-pAS | Short | WT | TAATACGACTCACTATAGGG | GCGAGGCTGTCTGCCCCCTCCCCTGCCAGCCCTACTCCCTAGTCTGCCCCCATCACAGTTCCCTATAGTGAGTCGTATTA |
| | | G/A-mutant | TAATACGACTCACTATAGGG | GCGAGGCTGTCTGCTCTCTCCTCTGCCAGCTCTACTCTCTAGTCTGCTCCCATCACAGTTCCCTATAGTGAGTCGTATTA |

| | | | | |
|---|---|---|---|---|
| | Long | WT | TAATACGACTCACTATAGGGAAAATAAATATAATTGGG ATTTTAAAAACATGAGAAATGTGAACTGTGATGGGGGC AGACTAGGGAGTAG | AAGGATGGGACCGGCAGCCAGGGATGAAGGGTGCGAGGCG AGGCTGTCTGCCCCCTCCCCTGCCAGCCCTACTCCCTAGT CTGCCCCC |
| | | G/A-mutant | TAATACGACTCACTATAGGGAAAATAAATATAATTGGG ATTTTAAAAACATGAGAAATGTGAACTGTGATGGGaGC AGACTAGaGAGTAG | AAGGATGGGACCGGCAGCCAGGGATGAAGGGTGCGAGGCG AGGCTGTCTGCTCTCTCCTCTGCCAGCTCTACTCTCTAGT CTGCTCCC |
| | | C/A-mutant | TAATACGACTCACTATAGGGAAAATAAATATAATTGGG ATTTTAAAAACATGAGAAATGTGAACTGTGATGGGGGC AGACTAGGGAGTAG | AAGGATGTGACCGGCAGCCAGTGATGAAGTGTGCGATGCG ATGCTGTCTGCCCCCTCCCCTGCCAGCCCTACTCCCTAGT CTGCCCCC |
| | | GC/AA-mutant | TAATACGACTCACTATAGGGAAAATAAATATAATTGGG ATTTTAAAAACATGAGAAATGTGAACTGTGATGGGAGC AGACTAGAGAGTAG | AAGGATGTGACCGGCAGCCAGTGATGAAGTGTGCGATGCG ATGCTGTCTGCTCTCTCCTCTGCCAGCTCTACTCTCTAGT CTGCTCCC |
| MAPK3 | Long | WT | TAATACGACTCACTATAGGGCGGGTGACAGGCAGGCGG GAAGGGGCG | CTCCACTCCTCCCCTCCCACCGCCCTCCTCCCCACGGCGG CCCCGCCCGAGGCCCCGCCCCTTCCCGCCTGCCTG |
| | | G/A-mutant | TAATACGACTCACTATAGGGCGGGTGACAGGCAGGCGG GAAGGGGCG | CTCCACTCCTCCCCTCCCACCGCCCTCCTCCCCACGGCGG CCCCGCCCGAGGCCCCGCCCCTTCCCGCCTGCCTG |
| SYNCRIP | Long | WT | TAATACGACTCACTATAGGGAGCTGGAGGAGGGCAGGG GCTGAGGGAGTGAGTGAAGCGGACGCGCGAG | CTCCGAGCGCGCTCCCGGTGCGCGCGCGCGCCCCCGCGTG ACCCCCCCTTCCCTTCCCTTCCCTTCCCTCCCCTCCCTCG CGCGTCCGCTTCACTC |
| | | G/A-mutant | TAATACGACTCACTATAGGGAGCTGGAGGAGGGCAGGG GCTGAGGGAGTGAGTGAAGCGGACGCGCGAG | CTCCGAGCGCGCTCTCGGTGCGCGCGCGCGCTCTCGCGTG ACTCTCTCTTCTCTTCTCTTCTCTTCTCTCTCCTCTCTCG CGCGTCCGCTTCACTC |
| PPP1CA | Long | WT | TAATACGACTCACTATAGGGCGGGGCCGCGGGCCGGGG GCGGACTGGG | CCTCCCGCCCTCCGGCAGCCTCCTTCCGGCCTGGCTCTCC TTCCGCCCGCCCCAGTCCGCCCCCGGCCCG |
| | | G/A-mutant | TAATACGACTCACTATAGGGCGAGGCCGCGAGCCGAGA GCGGACTGGA | CCTCCCGCCCTCCGGCAGCCTCCTTCCGGCCTGGCTCTCC TTCCGCCCGCCCCAGTCCGCCCCCGGCCCG |
| ERCC2 | Long | WT | TAATACGACTCACTATAGGGCGGGGGGTCTTGAAGATG GGGTCATCGGTGGGCGCGCCTG | TCAGTGTCTCCTCGCTATCACTGCTGCTCCCTGCGGCTGC CCCCGTCCCACCCCTTCACCCTCCCCTCGCCCCCTTGGGG ACCCAGGCGCGCCCACCGAT |
| | | G/A-mutant | TAATACGACTCACTATAGGGCGGAGGATCTTGAAGATG AGGTCATCGGTGAGCGCGCCTG | TCAGTGTCTCCTCGCTATCACTGCTGCTCCCTGCGGCTGC TCTCGTCTCACTCCTTCACTCTCCTCTCGCTCTCTTGGGG ACTCAGGCGCGCTCACCGAT |
| ESR2 | Long | WT | TAATACGACTCACTATAGGGAGTGTCAGAGCTGGAGCG CGCGTGGCCCCCTCTGTGTTGG | CAGACCTGCTGGGGGGTGGGGACGTGCGGGTGACAAAATC CAGACTACGACCCTCCCTGAGCCCTGGCAACCCCGGGGTG ACCCCAACACAGAGGGGGCC |

| | | | | |
|---|---|---|---|---|
| | | G/A-mutant | TAATACGACTCACTATAGGGAGTGTCAGAGCTGGAGCGCGCGTGGCCCCCTCTGTGTTGG | CAGACCTGCTGGGGGGTGGGGACGTGCGGGTGACAAAATCCAGACTACGACTCTCTCTGAGCTCTGGCAACTCCGGGGTGACTCCAACACAGAGGGGGCC |
| TCF7L1 | Long | WT | TAATACGACTCACTATAGGGCGCCGGGCCGGGCCGGGCAGGGCGCGGGCGGCTAGGGGCTCCGAGAGCGGCGGCCC | GGTGGGGCCGCGGGCCGGGGCCGCCGCTCTCGGAGCC |
| | | G/A-mutant | TAATACGACTCACTATAGGGCGCCGAGCCGAGCCGAGCAGAGCGCGAGCGGCTAGAGGCTCCGAGAGCGGCGGCCC | GGTGGGGCCGCGGGCCGGGGCCGCCGCTCTCGGAGCC |
| SMAD2 | Long | WT | TAATACGACTCACTATAGGGCGCCCGGGCCGCCGGCCGGGCCCGGGCCTGGGGGCGGGGCGGGAAGACGGCGGCCGGGAGTG | GGGAATGGGCGATTGGAGGCGGAACTGAAAACACTCCCGGCCGCCGTCTTC |
| | | G/A-mutant | TAATACGACTCACTATAGGGCGCCCGAGCCGCCGGCCGAGCCCGAGCCTGAGAGCGGAGCGAGAAGACGGCGGCCGGGAGTG | GGGAATGGGCGATTGGAGGCGGAACTGAAAACACTCCCGGCCGCCGTCTTC |
| SMAD7 | Long | WT | TAATACGACTCACTATAGGGCGGAGAGCCGCGCAGGGCGCGGGCCGCGCGGGGTGGGGCAGCCGGAGCGCAGGCCCC | GCGGGGGCCCGGGGGCGCCCGCCGGGGATCGGGGGCCTGCGCTCCGGCTGC |
| | | G/A-mutant | TAATACGACTCACTATAGGGCGGAGAGCCGCGCAGAGCGCGAGCCGCGCGGAGTAGAGCAGCCGGAGCGCAGGCCCC | GCGGGGGCCCGGGGGCGCCCGCCGGGGATCGGGGGCCTGCGCTCCGGCTGC |
| DNMT3B | Long | WT | TAATACGACTCACTATAGGGAGGAGGGAAAATAATGCACTGGCTTCCTGAGCCCCTGCAGAGGCTG | AATGCCATTTAGGACCTGCCTGTCCCGGCCCGTCTGCCGCCAGCCTCTGTCCCCTCTGGCCCTGGCCCCCTCTCCCTGCTCAGCCTCTGCAGGGGCTCAG |
| | | G/A-mutant | TAATACGACTCACTATAGGGAGGAGGGAAAATAATGCACTGGCTTCCTGAGCCCCTGCAGAGGCTG | AATGCCATTTAGGACCTGCCTGTCCCGGCTCGTCTGCCGCCAGCCTCTGTCTCCTCTGGCTCTGGCTCTCTCTCTCTGCTCAGCCTCTGCAGGGGCTCAG |
| CREM | Long | WT | TAATACGACTCACTATAGGGAGCCTGGATTTTTTTCCTCGGGGCCTCCCCCGGGAGGCCGTC | CTGCCGACCCCGGACCAAAAGTAGCGCTGCAGCCGACCGAACCGCGTCCTCCCGCCCCGTCCTCCCCTCCCCCACGCCGGGACGGCCTCCCGGGGGAGGC |
| | | G/A-mutant | TAATACGACTCACTATAGGGAGCCTGGATTTTTTTCCTCGGGGCCTCCCCCGGGAGGCCGTC | CTGCCGACCCCGGACCAAAAGTAGCGCTGCAGCCGACCGAACCGCGTCCTCTCGCTCCGTCCTCCTCTCTCTCACGCCGGGACGGCCTCCCGGGGGAGGC |
| ACVR1C | Long | WT | TAATACGACTCACTATAGGGCCGCCCGGCTGCGGGGCCAGTGGCAGGAGCGCCGCGCACCGCCAGCC | ACACCCTTTTGAAGTGCGCGGTTGGCTCTAGTCAGTGTGGGCGCCCCCCTCCCCGGCCGCCCCCATCCCACGCCCCCTGCGGCTGGCGGTGCGCGGCGCT |

| | | | | |
|---|---|---|---|---|
| | | G/A-mutant | TAATACGACTCACTATAGGGCCGCCCGGCTGCGGGGCCAGTGGCAGGAGCGCCGCGCACCGCCAGCC | ACACCCTTTTGAAGTGCGCGGTTGGCTCTAGTCAGTGTGGGCGCTCTCCTCTCCGGCCGCTCTCATCTCACGCTCTCTGCGGCTGGCGGTGCGCGGCGCT |
| GNAI2 | Long | WT | TAATACGACTCACTATAGGGCCGACCCGAGTGCTTCCCGCAGAGGGCTGGTGGTGGG | CCCGCCGTCCGCCGGCCCGGCCGCCGCCCGGCCCCCACACGGCCCACGGCCCGGCTCGGCCCCGCCCGACCCACTCCGCTCCCACCACCAGCCCTCTGCG |
| | | G/A-mutant | TAATACGACTCACTATAGGGCCGACCCGAGTGCTTCCCGCAGAGGGCTGGTGGTGGG | CCCGCCGTCCGCCGGCCCGGCCGCCGCCCGGCCCCCACACGGCTCACGGCTCGGCTCGGCTCCGCTCGACCCACTCCGCTCCCACCACCAGCCCTCTGCG |
| PTPRJ | Long | WT | TAATACGACTCACTATAGGGCTAGGCTCCGGCGTGTGGCCGCGGCCGCCGCCGCCGCTGCCATGTC | CCTCCTGCCGTCTCCTTCGCCTCCTCCTCCGCCAGCCGGTCCGCCTCGTCCCCGCTCCGCCCGCCCCGGGCTTCCCCGGAGACATGGCAGCGGCGGCGGC |
| | | G/A-mutant | TAATACGACTCACTATAGGGCTAGGCTCCGGCGTGTGGCCGCGGCCGCCGCCGCCGCTGCCATGTC | CCTCCTGCCGTCTCCTTCGCCTCCTCCTCCGCCAGCCGGTCCGCCTCGTCCTCGCTCCGCTCGCCTCGGGCTTCTCCGGAGACATGGCAGCGGCGGCGGC |
| MYCL1 | Long | WT | TAATACGACTCACTATAGGGAGCCGGTCCGCTCCAGGTGGCGGGCGGCTGGAGCGAGGTGA | GGGGCGCGCCGTGCCCAGAAGGCAGCCTGCAGCCAGCCCGCACCGCGGGACCCGCGCCCGTGCCCTGGCCACCCGCAGCCTCACCTCGCTCCAGCCGCCC |
| | | G/A-mutant | TAATACGACTCACTATAGGGAGCCGGTCCGCTCCAGGTGGCGGGCGGCTGGAGCGAGGTGA | GGGGCGCGCCGTGCCCAGAAGGCAGCCTGCAGCCAGCCCGCACCGCGGGACTCGCGCTCGTGCTCTGGCCACTCGCAGCCTCACCTCGCTCCAGCCGCCC |

**Supplementary table S2** RNA sequences of the PG4 probed *in vitro* by in-line probing

| Figure | Candidate | | Sequences 5'-3' | |
|---|---|---|---|---|
| **S1** | NCAM2 | Short | WT | GGAGGAGCGCGCGGGCUGCGGGCGGCUGGGGCACCGCGGGAGCGGCGGCGGCGG |
| | | | G/A-mutant | GGAGGAGCGCGCGAGCUGCGAGCGGCUGAAGCACCGCGAGAGCGGCGGCGGCGG |
| | | Long | WT | GGGAUAGUGCGGCAAGAGCGGAGCUUGCAGUCACUUUGCGAGGAGGAGCGCGCGGGCUGCGGGCGGCUGGGGCACC GCGGGAGCGGCGGCGGCGGCUCUAGCAGAGGCGGCCGGGGCAGCGAAAGGUUCUC |
| | | | G/A-mutant | GGGAUAGUGCGGCAAGAGCGGAGCUUGCAGUCACUUUGCGAGGAGGAGCGCGCGAGCUGCGAGCGGCUGAAGCACC GCGAGAGCGGCGGCGGCGGCUCUAGCAGAGGCGGCCGGGGCAGCGAAAGGUUCUC |
| **S2** | BARHL1 | Short | WT | GGGGCAGCGGCGGCUGGGGUUGGGGGUGGGUGGGGAGCUUUUGGGG |
| | | | G/A-mutant | GGGGCAGCGGCGGCUGAAGUUGAGAGUGGGUGAAGAGCUUUUGGGG |
| | | Long | WT | GGGCCCGCCUUCCCCAUGCCAGCCCGCAGCUAGGGGCAGGGGCAGCGGCGGCUGGGGUUGGGGGUGGGUGGGGAGC UUUUGGGGAGGACAGGUCGCAGCUUGGCUGCUAGCCACCAUGACUUC |
| | | | G/A-mutant | GGGCCCGCCUUCCCCAUGCCAGCCCGCAGCUAGGGGCAGGGGCAGCGGCGGCUGAAGUUGAGAGUGGGUGAAGAGC UUUUGGGGAGGACAGGUCGCAGCUUGGCUGCUAGCCACCAUGACUUC |
| **S3** | FZD2 | Short | WT | GGGGAAGAAGCGCAGUCUCCGGGUUGGGGGCGGGGGCGGGGGGGGCGCCAAGGAGCCGGG |
| | | | G/A-mutant | GGGGAAGAAGCGCAGUCUCCGGGUUGGAAGCGAGAGCGAAGGAAGCGCCAAGGAGCCGGG |
| | | Long | WT | GGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGUCUCCGGGUUGGGGCGGGGGCGGGGGGGGCG CCAAGGAGCCGGGUGGGGGGCGGCGGCCAGCGCUAGCCACCAUGACU |
| | | | G/A-mutant | GGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGUCUCCGGGUUGGAAGCGAGAGCGAAGGAAGC GCCAAGGAGCCGGGUGGGGGGCGGCGGCCAGCGCUAGCCACCAUGACU |
| **S4** | EBAG9 | Short | WT | GGAGCCUCCGCCGGGCGGGCGGGGAGGGGGAGGGGCAGGUUUUGA |
| | | | G/A-mutant | GGAGCCUCCGCCGGGCAAGCAAGGAGAAGGAGGGGCAGGUUUUGA |
| | | Long | WT | GGGAGCGCGCCUUGUGUGCGCGCGCGGCCCGCGGCAGCUCGGAGCCUCCGCCGGGCGGGCGGGGAGGGGGAGGGGC AGGUUUUGAUUCCCACCGCUAGCCACCAUGACUUCGAAAGUUUAUGAUC |
| | | | G/A-mutant | GGGAGCGCGCCUUGUGUGCGCGCGCGGCCCGCGGCAGCUCGGAGCCUCCGCCGGGCAAGCAAGGAGAAGGAGGGGC AGGUUUUGAUUCCCACCGCUAGCCACCAUGACUUCGAAAGUUUAUGAUC |
| **S5** | FXR1 | Short | WT | GGGCAAAAUAUGGCAGAGAUGGGGGGUGGUGGGUGGGGUGGGAUCCCUAACGUAAUA |
| | | | G/A-mutant | GGGCAAAAUAUGGCAGAGAUGGAGAGUGGUGAGUGAGGUGAGAUCCCUAACGUAAUA |

| | | | | |
|---|---|---|---|---|
| | | Long | WT | GGGUUGCUGGCUAUAGGAAAUGUUAUUUUGUUUUCAAAAUAUGGCAGAGAUGGGGGGUGGUGGGUGGGGUGGGAUC<br>CCUAACGUAAUAUUCUUUAUGAAAGCAUUAGCUGCUUUUGUUACAUU |
| | | | G/A-mutant | GGGUUGCUGGCUAUAGGAAAUGUUAUUUUGUUUUCAAAAUAUGGCAGAGAUGGAGAGUGGUGAGUGAGGUGAGAUC<br>CCUAACGUAAUAUUCUUUAUGAAAGCAUUAGCUGCUUUUGUUAGAUU |
| **S6** | LRP5 | Short | WT | GGGAACUGUGAUGGGGUGGGCAGGGCUGGGAGAACUUUGUACAG |
| | | | G/A-mutant | GGGAACUGUGAUGGAGUGAGCAGAGCUGAGAGAACUUUGUACAG |
| | | Long | WT | GGGAAAAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUGAUGGGGUGGGCAGGGCUGGGAGAAC<br>UUUGUACAGUGGAGAAAUAUUUAUAAACUUAAUUUUGUAAAACAG |
| | | | G/A-mutant | GGGAAAAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUGAUGGAGUGAGCAGAGCUGAGAGAAC<br>UUUGUACAGUGGAGAAAUAUUUAUAAACUUAAUUUUGUAAAACAG |
| **S7** | AASDHPPT | Short | WT | GGGGGCGGGCUGGGAGGGCUGUCGGUGGGCCAGUCUGC |
| | | | G/A-mutant | GAGAGCGAGCUGAGAGAGCUGUCGGUGGGCCAGUCUGC |
| | | Long | WT | GGGAUUACAGCUCUUAAGGCUAGAGUACUUAAUACGACUCACUAUAGGCUAGCGGGGGCGGGCUGGGAGGGCUGUC<br>GGUGGGCCAGUCUGCGUAGCGACGGCCCGUCCCCUGCGCACGGAC |
| | | | G/A-mutant | GGGAUUACAGCUCUUAAGGCUAGAGUACUUAAUACGACUCACUAUAGGCUAGCGAGAGCGAGCUGAGAGAGCUGUC<br>GGUGGGCCAGUCUGCGUAGCGACGGCCCGUCCCCUGCGCACGGAC |
| **S8** | THRA1 | Short | WT | GGGUGCUGUGCCCUAGGGCCUGGGUGGCAGGGGGUGGGUGGCCUGUGGG |
| | | | G/A-mutant | GGGUGCUGUGCCCGAGGGCCUGGGUGGCAAAAAAUAAAUGGCCUGUGGG |
| | | Long | WT | GGGUGGUGUGAAAGGCCAAGUGCUGAGGCGGGUAUCAUGGGUGCUGUGCCCUAGGGCCUGGGUGGCAGGGGGUGGG<br>UGGCCUGUGGGUGUGCCGGGGGGGCCAGUGUGCCCACCCCAGUCUCUUGG |
| | | | G/A-mutant | GGGUGGUGUGAAAGGCCAAGUGCUGAGGCGGGUAUCAUGGGUGCUGUGCCCUAGGGCCUGGGUGGCAUAAAAUAAU<br>UGGCCUGUGGGUGUGCCGGGGGGGCCAGUGUGCCCACCCCAGUCUCUUGG |
| **S9** | DOC2B | Short | WT | GGGCCUGCUGAGCCGCCCCCGGGCCGGGGUCGCGCCGGGCCGGGCCGCGCCCGGGG |
| | | | G/A-mutant | GGGCCUGCUGAGCCGCCCCCGAGCCGAAGUCGAGCCGAGCCGAGCCGCGCCCGGGG |
| | | | C/A-mutant | GGGAAUGAUGAGAAGCCCACGGGACGGGGUCGAGAAGGGCCGGGACGCGCACGGGG |
| | | | GC/AA-mutant | GGGAAUGAUGAGAAGCCCACGAGACGAAGUCGAGAAGAGCCGAGACGCGCACGGGG |
| | | Long | WT | GGGCGAUGCCCGCAGCCCCCGCCGCGCCCCGCCGGGCCUGCUGAGCCGCCCCCGGGCCGGGGUCGCGCCGGGCCGG<br>GCCGCGCCCGGGGCGGGGCGGCGCUGCCUGCGCUAGCCACCAUGACUUCGA |

| | | | | |
|---|---|---|---|---|
| | | | G/A-mutant | GGGCGAUGCCCGCAGCCCCCGCCGCGCCCCGCCGGGCCUGCUGAGCCGCCCCCGAGCCGAAGUCGCGCCGAGCCGAGCCGCGCCCGGGGCGGGGCGGCGCUGCCUGCGCUAGCCACCAUGACUUCGA |
| | | | C/A-mutant | GGGCGAUGCCCGCAGCCCCCGCCGCGCCCCGCCGGGAAUGUUGAGAAGCCCACGGGACGGGGUCGAGAAGGGCCGGGACGCGCACGGGGCGGGGCGGCGCUGCCUGCGCUAGCCACCAUGACUUCGA |
| | | | GC/AA-mutant | GGGCGAUGCCCGCAGCCCCCGCCGCGCCCCGCCGGGAAUGUUGAGAAGCCCACGAGACGAAGUCGAGAAGAGCCGAGACGCGCACGGGGCGGGGCGGCGCUGCCUGCGCUAGCCACCAUGACUUCGA |
| **S10** | TNSF12 | Short | WT | GGCUCCCCCUCCCCCGAUCCCUCGGGUCCCGGGAUGGGGGGGCGGUGAGGCAGG |
| | | | G/A-mutant | GGCUCCCCCUCCCCCGAUCCCUCGAGUCCCGAGAUGAGGAAGCGGUGAGGCAGG |
| | | | C/A-mutant | GGAUCACACUCACACGAUAAAUAGGGUCCCGGGAUGGGGGGGAGGUGAGGCAGG |
| | | | GC/AA-mutant | GGAUCACACUCACACGAUAAAUAGAGUCCCGAGAUGAGGAAGAGGUGAGGCAGG |
| | | Long | WT | GGGCCUCUCCCCGGCCCGAUCCGCCCGCCGGCUCCCCCUCCCCCGAUCCCUCGGGUCCCGGGAUGGGGGGGCGGUGAGGCAGGCACAGCCCCCCGCCCCCGCUAGCCACCAUGACUUCGAA |
| | | | G/A-mutant | GGGCCUCUCCCCGGCCCGAUCCGCCCGCCGGCUCCCCCUCCCCCGAUCCCUCGAGUCCCGAGAUGAGGAAGCGGUGAGGCAGGCACAGCCCCCCGCCCCCGCUAGCCACCAUGACUUCGAA |
| | | | C/A-mutant | GGGCCUCUCCCCGGCCCGAUCCGCCCGCCGGAUCACACUCACACGAUAAAUAGGGUCCCGGGAUGGGGGGGAGGUGAGGCAGGCACAGCCCCCCGCCCCCGCUAGCCACCAUGACUUCGAA |
| | | | GC/AA-mutant | GGGCCUCUCCCCGGCCCGAUCCGCCCGCCGGAUCACACUCACACGAUAAAUAGAGUCCCGAGAUGAGGAAGAGGUGAGGCAGGCACAGCCCCCCGCCCCCGCUAGCCACCAUGACUUCGAA |
| **S11** | MAP3K11 | Short | WT | GGCUCCCCAGAGAGGCGUGGGUCUGGGGCUGAGGGCCAGGGCCCGGAUGCCCAGG |
| | | | G/A-mutant | GGGCUCCCCAGAGAGGCGUGAGUCUGGAGCUGAGAGCCAGAGCCCGGAUGCCCAGG |
| | | | C/A-mutant | GGAUCACCAGAGAGGCGUGGGUCUGGGGAUGAGGGACAGGGACCGGAUGCCAAGG |
| | | | GC/AA-mutant | GGAUCACCAGAGAGGCGUGAGUCUGGAGAUGAGAGACAGAGACCGGAUGCCAAGG |
| | | Long | WT | GGGCGAGAUGCGGGGGGCCGGGAGACAACACUCCUGGCUCCCCAGAGAGGCGUGGGUCUGGGGCUGAGGGCCAGGGCCCGGAUGCCCAGGUUCCGGGACUAGGGCCUUGGCAGCCAGCGGGGGUGG |
| | | | G/A-mutant | GGGCGAGAUGCGGGGGGCCGGGAGACAACACUCCUGGCUCCCCAGAGAGGCGUGAGUCUGGAGCUGAGAGCCAGAGCCCGGAUGCCCAGGUUCCGGGACUAGGGCCUUGGCAGCCAGCGGGGGUGG |
| | | | C/A-mutant | GGGCGAGAUGCGGGGGGCCGGGAGACAACACUCCUGGAUCACCAGAGAGGCGUGGGUCUGGGGUUGAGGGACAGGGACCGGAUGCCAAGGUUCCGGGACUAGGGCCUUGGCAGCCAGCGGGGGUGG |

| | | | | |
|---|---|---|---|---|
| | | | GC/AA-mutant | GGGCGAGAUGCGGGGGGCCGGGAGACAACACUCCUGGAUCACCAGAGAGGCGUGAGUCUGGAGUUGAGAGACAGAG ACCGGAUGCCAAGGUUCCGGGACUAGGGCCUUGGCAGCCAGCGGGGGUGG |
| **S12** | TTYH1 | Short | WT | GGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGGCAGACAGCCUCGC |
| | | | G/A-mutant | GGGAGUAGCUGAGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAGAGCAGACAGCCUCGC |
| | | Long | WT | GGGUGCUCCCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGG GGCAGACAGCCUCGCCUCGCACCCUUCAUCCCUGGCUGCCGGUCCCAUCCUU |
| | | | G/A-mutant | GGGUGCUCCCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAG AGCAGACAGCCUCGCCUCGCACCCUUCAUCCCUGGCUGCCGGUCCCAUCCUU |
| | | | C/A-mutant-1 | GGGUGCUCCCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGG GGCAGACAGCAUGCGAUCGCACACUUCAUCACUGGCUGCCGGUCACAUCCUU |
| | | | GC/AA-mutant-1 | GGGUGCUCCCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAG AGCAGACAGCAUGCGAUCGCACACUUCAUCACUGGCUGCCGGUCACAUCCUU |
| | | | C/A-mutant-2 | GGGUGAUCACAUUUCUGUACUUGGAAUUGGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGG GGGCAGACAGCAUCGCAUCGCACACUUCAUCACUGGCUGCCGGUCACAUCCUU |
| | | | GC/AA-mutant-2 | GGGUGAUCACAUUUCUGUACUUGGAAUUGGGAGUAGCUGAGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAG AGCAGACAGCAUCGCAUCGCACACUUCAUCACUGGCUGCCGGUCACAUCCUU |
| **S13** | MAPK3 | Long | WT | GGGCGGGUGACAGGCAGGCGGGAAGGGGCGGGGCCUCGGGCGGGGCCGCCGUGGGGAGGAGGGCGGUGGGAGGGGA GGAGUGGAG |
| | | | G/A-mutant | GGGCGGGUGACAGGCAGGCGAGAAGAGACGGAGCCUCGAGCGAGGCCGCCGUGAGGAGGAGAGCGGUGAGAGGAGA GGAGUGGAG |
| **S14** | SYNCRIP | Long | WT | GGGAGCUGGAGGAGGGCAGGGGCUGAGGGAGUGAGUGAAGCGGACGCGCGAGGGAGGGGAGGGAAGGGAAGGGAAG GGAAGGGGGGGUCACGCGGGGGCGCGCGCGCGCACCGGGAGCGCGCUCGGAG |
| | | | G/A-mutant | GGGAGCUGGAGGAGGGCAGGGCUGAGGGAGUGAGUGAAGCGGACGCGCGAGAGAGGAGAGAGAAGAGAAGAGAAGA GAAGAGAGAGUCACGCGAGAGCGCGCGCGCGCACCGAGAGCGCGCUCGGAG |
| **S15** | PPP1CA | Long | WT | GGGCGGGGCCGCGGGCCGGGGGCGGACUGGGGCGGGCGGAAGGAGAGCCAGGCCGGAAGGAGGCUGCCGGAGGGCG GGAGG |
| | | | G/A-mutant | GGGCGAGGCCGCGAGCCGAGAGCGGACUGGAGCGGGCGGAAGGAGAGCCAGGCCGGAAGGAGGCUGCCGGAGGGCG GGACC |
| **S16** | ERCC2 | Long | WT | GGGCGGGGGGUCUUGAAGAUGGGGUCAUCGGUGGGCGCGCCUGGGUCCCCAAGGGGGCGAGGGGAGGGUGAAGGGG UGGGACGGGGGCAGCCGCAGGGAGCAGCAGUGAUAGCGAGGAGACACUGA |

| | | | | |
|---|---|---|---|---|
| | | | G/A-mutant | GGGCGGAGGAUCUUGAAGAUGAGGUCAUCGGUGAGCGCGCCUGAGUCCCCAAGAGAGCGAGAGGAGAGUGAAGGAGUGAGACGAGAGCAGCCGCAGGGAGCAGCAGUGAUAGCGAGGAGACACUGA |
| **S17** | ESR2 | Long | WT | GGGAGUGUCAGAGCUGGAGCGCGCGUGGCCCCCUCUGUGUUGGGGUCACCCCGGGGUUGCCAGGGCUCAGGGAGGGUCGUAGUCUGGAUUUUCUCACCCGCACGUCCCCACCCCCCAGCAGGUCUG |
| | | | G/A-mutant | GGGAGUGUCAGAGCUGGAGCGCGCGUGGCCCCCUCUGUGUUCCACUCACCCCGGAGUUGCCAGAGCUCAGAGAGAUCGUAGUCUGGAUUUUGUCACCCGCACGUCCCCACCCCCCAGCAGGUCUG |
| **S18** | TCF7L1 | Long | WT | GGGCGCCGGGCCGGGCCGGGCAGGGCGCGGGCGGCUAGGGGCUCCGAGAGCGGCGGCCCCGGCCCGCGGCCCACC |
| | | | G/A-mutant | GGGCGCCGAGCCGAGCCGAGCAGAGCGCGAGCGGCUAGAGGCUCCGAGAGCGGCGGCCCCGGCCCGCGGCCCCACC |
| **S19** | SMAD2 | Long | WT | GGGCGCCCGGGCCGCCGGCCGGGCCCGGGCCUGGGGGCGGGGCGGGAAGACGGCGGCCGGGAGUGUUUUCAGUUCCGCCUCCAAUCGCCCAUUCCC |
| | | | G/A-mutant | GGGCGCCCGAGCCGCCGGCCGAGCCCGAGCCUGAGAGCGGAGCGAGAAGACGGCGGCCGGGAGUGUUUUCAGUUCCGCCUCCAAUCGCCCAUUCCC |
| **S20** | SMAD7 | Long | WT | GGGCGGAGAGCCGCGCAGGGCGCGGGCCGCGCGGGGUGGGGCAGCCGGAGCGCAGGCCCCCGAUCCCCGGCGGGCGCCCCCGGGCCCCCGC |
| | | | G/A-mutant | GGGCGGAGAGCCGCGCAGAGCGCGAGCCGCGCGGAGUAGAGCAGCCGGAGCGCAGGCCCCCGAUCCCCGGCGGGCGCCCCCGGGCCCCCGC |
| **S21** | DNMT3B | Long | WT | GGGAGGAGGGAAAAUAAUGCACUGGCUUCCUGAGCCCCUGCAGAGGCUGAGCAGGGAGAGGGGGCCAGGGCCAGAGGGGACAGAGGCUGGCGGCAGACGGGCCGGGACAGGCAGGUCCUAAAUGGCAUU |
| | | | G/A-mutant | GGGAGGAGGGAAAAUAAUGCACUGGCUUCCUGAGCCCCUGCAGAGGCUGAGCAGAGAGAGAGAGCCAGAGCCAGAGGAGACAGAGGCUGGCGGCAGACGAGCCGGGACAGGCAGGUCCUGGGUGGCAUU |
| **S22** | CREM | Long | WT | GGGAGCCUGGAUUUUUUUCCUCGGGGCCUCCCCCGGGAGGCCGUCCCGGCGUGGGGGAGGGGAGGACGGGGCGGGAGGACGCGGUUCGGUCGGCUGCAGCGCUACUUUUGGUCCGGGGUCGGCAG |
| | | | G/A-mutant | GGGAGCCUGGAUUUUUUUCCUCGGGGCCUCCCCCGGGAGGCCGUCCCGGCGUGAGAGAGAGGAGGACGGAGCGAGAGGACGCGGUUCGGUCGGCUGCAGCGCUACUUUUGGUCCGGGGUCGGCAG |
| **S23** | ACVR1C | Long | WT | GGGCCGCCCGGCUGCGGGGCCAGUGGCAGGAGCGCCGCGCACCGCCAGCCGCAGGGGGCGUGGGAUGGGGGCGGCCGGGGAGGGGGGCGCCCACACUGACUAGAGCCAACCGCGCACUUCAAAAGGGUGU |
| | | | G/A-mutant | GGGCCGCCCGGCUGCGGGGCCAGUGGCAGGAGCGCCGCGCACCGCCAGCCGCAGAGAGCGUGAGAUGAGAGCGGCCGGAGAGGAGAGCGCCCACAGUGACUAGAGCCAACCGCGCACUUCAAAAGGGUGU |
| **S24** | GNAI2 | Long | WT | GGGCCGACCCGAGUGCUUCCCGCAGAGGGCUGGUGGUGGGAGCGGAGUGGGUCGGGCGGGGCCGAGCCGGGCCGUGGGCCGUGUGGGGGCCGGGCGGCGGCCGGGCCGGCGGACGGCGGG |
| | | | G/A-mutant | GGGCCGACCCGAGUGCUUCCCGCAGAGGGCUGGUGGUGGGAGCGGAGUGGGUCGAGCGGAGCCGAGCCGAGCCGUGAGCCGUGUGGGGGCCGGGCGGCGGCCGGGCCGGCGGACGGCGGG |
| **S25** | PTPRJ | Long | WT | GGGCUAGGCUCCGGCGUGUGGCCGCGGCCGCCGCCGCCGCUGCCAUGUCUCCGGGGAAGCCCGGGGCGGGCGGAGCGGGGACGAGGCGGACCGGCUGGCGGAGGAGGAGGCGAAGGAGACGGCAGGAGG |

| | | | | |
|---|---|---|---|---|
| | | | G/A-mutant | GGGCUAGGCUCCGGCGUGUGGCCGCGGCCGCCGCCGCCGCUGCCAUGUCUCCGGAGAAGCCCGAGGCGAGCGGAGC<br>GAGGACGAGGCGGACCGGCUGGCGGAGGAGGAGGCGAAGGAGACGGCAGGAGG |
| **S26** | MYCL1 | Long | WT | GGGAGCCGGUCCGCUCCAGGUGGCGGGCGGCUGGAGCGAGGUGAGGCUGCGGGUGGCCAGGGCACGGGCGCGGGUC<br>CCGCGGUGCGGGCUGGCUGCAGGCUGCCUUCUGGGCACGGCGCGCCCC |
| | | | G/A-mutant | GGGAGCCGGUCCGCUCCAGGUGGCGGGCGGCUGGAGCGAGGUGAGGCUGCGAGUGGCCAGAGCACGAGCGCGAGUC<br>CCGCGGUGCGGGCUGGCUGCAGGCUGCCUUCUGGGCACGGCGCGCCCC |

**Supplementary table S3** Full length 3'-UTR RNA sequences of candidates

| **Constructions** | **Full length 3'-UTR (5' to 3')** | |
|---|---|---|
| **LRP5** | WT | CCUCGGCCGGGCCACUCUGGCUUCUCUGUGCCCCUGUAAAUAGUUUUAAAUAUGAACAAAGAAAAAAAUAUAUUUUAUGAUUUAAAA<br>AAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUGAUGGGGUGGGCAGGGCUGGGAGAACUUUGUACAGUGGAGAA<br>AUAUUUAUAAACUUAAUUUUGUAAAACAG |
| | G/A-mutant | CCUCGGCCGGGCCACUCUGGCUUCUCUGUGCCCCUGUAAAUAGUUUUAAAUAUGAACAAAGAAAAAAAUAUAUUUUAUGAUUUAAAA<br>AAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUGAUGGAGUGAGCAGAGCUGAGAGAACUUUGUACAGUGGAGAA<br>AUAUUUAUAAACUUAAUUUUGUAAAACAG |
| **TTYH1** | WT | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCUCUUGGCCACUGUGCUC<br>CCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGGCAGACAGCCUCGCCUCG<br>CACCCUUCAUCCCUGGCUGCCGGUCCCAUCCUUGGAGGGACUAAGCUGGGGGUGGGGGGACAUGAGUCCCCCUGCUGCCCCUGCCACA<br>UCCCAGUGGGCUCUGACCCCCUGAUCUCAACUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAAAUAAAAGGGAAGACUAUUUUAC |
| | G/A-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCUCUUGGCCACUGUGCUC<br>CCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAGAGCAGACAGCCUCGCCUCG<br>CACCCUUCAUCCCUGGCUGCCGGUCCCAUCCUUGGAGGGACUAAGCUGGGGGUGGGGGGACAUGAGUCCCCCUGCUGCCCCUGCCACA<br>UCCCAGUGGGCUCUGACCCCCUGAUCUCAACUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAAAUAAAAGGGAAGACUAUUUUAC |
| | C/A-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCUCUUGGCCACUGUGCUC<br>CCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGGCAGACAGCAUCGCAUCG<br>CACACUUCAUCACUGGCUGCCGGUCACAUCCUUGGAGGGACUAAGCUGGGGGUGGGGGGACAUGAGUCCCCCUGCUGCCCCUGCCACA<br>UCCCAGUGGGCUCUGACCCCCUGAUCUCAACUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAAAUAAAAGGGAAGACUAUUUUAC |
| | GC/AA-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCUCUUGGCCACUGUGCUC<br>CCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAGAGCAGACAGCAUCGCAUCG<br>CACACUUCAUCACUGGCUGCCGGUCACAUCCUUGGAGGGACUAAGCUGGGGGUGGGGGGACAUGAGUCCCCCUGCUGCCCCUGCCACA<br>UCCCAGUGGGCUCUGACCCCCUGAUCUCAACUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAAAUAAAAGGGAAGACUAUUUUAC |

| | | |
|---|---|---|
| **TTYH1 + pAS** | WT | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCAAUAAAUCUUGGCCACU<br>GUGCUCCCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGGCAGACAGCCUC<br>GCCUCGCACCCUUCAUCCCUGGCUGCCGGUCCCAUCCUUGGAGGGACUAAGCUGGGGGUGGGGGACAUGAGUCCCCCUGCUGCCCCU<br>GCCACAUCCCAGUGGGCUCUGACCCCCUGAUCUCAACUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAACUAACAGGGAAGACUA<br>UUUUAC |
| | G/A-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCAAUAAAUCUUGGCCACU<br>GUGCUCCCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAGAGCAGACAGCCUC<br>GCCUCGCACCCUUCAUCCCUGGCUGCCGGUCCCAUCCUUGGAGGGACUAAGCUGGGGGUGGGGGACAUGAGUCCCCCUGCUGCCCCU<br>GCCACAUCCCAGUGGGCUCUGACCCCCUGAUCUCAACUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAACUAACAGGGAAGACUA<br>UUUUAC |
| | C/A-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCAAUAAAUCUUGGCCACU<br>GUGCUCCCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGGCAGACAGCAUC<br>GCAUCGCACACUUCAUCACUGGCUGCCGGUCACAUCCUUGGAGGGACUAAGCUGGGGGUGGGGGACAUGAGUCCCCCUGCUGCCCCU<br>GCCACAUCCCAGUGGGCUCUGACCCCCUGAUCUCAACUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAACUAACAGGGAAGACUA<br>UUUUAC |
| | GC/AA-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCAAUAAAUCUUGGCCACU<br>GUGCUCCCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAGAGCAGACAGCAUC<br>GCAUCGCACACUUCAUCACUGGCUGCCGGUCACAUCCUUGGAGGGACUAAGCUGGGGGUGGGGGACAUGAGUCCCCCUGCUGCCCCU<br>GCCACAUCCCAGUGGGCUCUGACCCCCUGAUCUCAACUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAACUAACAGGGAAGACUA<br>UUUUAC |
| **TTYH1-LRP5-pAS** | WT | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCUCUUGGCCACUGUGCUC<br>CCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGAUUUAAAAAAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUG<br>AUGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGGCAGACAGCCUCGCCUCGCACCCUUCAUCCCUGGCUGCCGGUCCCAUC<br>CUUGGAGGGACUAAGCUGGGGGUGGGGGACAUGAGUCCCCCUGCUGCCCCUGCCACAUCCCAGUGGGCUCUGACCCCCUGAUCUCAA<br>CUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAACUAACAGGGAAGACUAUUUUUAC |
| | G/A-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCUCUUGGCCACUGUGCUC<br>CCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGAUUUAAAAAAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUG<br>AUGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAGAGCAGACAGCCUCGCCUCGCACCCUUCAUCCCUGGCUGCCGGUCCCAUC<br>CUUGGAGGGACUAAGCUGGGGGUGGGGGACAUGAGUCCCCCUGCUGCCCCUGCCACAUCCCAGUGGGCUCUGACCCCCUGAUCUCAA<br>CUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAACUAACAGGGAAGACUAUUUUUAC |

| | | |
|---|---|---|
| | C/A-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCUCUUGGCCACUGUGCUC<br>CCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGAUUUAAAAAAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUG<br>AUGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGGCAGACAGCAUCGCAUCGCACACUUCAUCACUGGCUGCCGGUCACAUC<br>CUUGGAGGGACUAAGCUGGGGGUGGGGGACAUGAGUCCCCCUGCUGCCCCUGCCACAUCCCAGUGGGCUCUGACCCCCUGAUCUCAA<br>CUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAACUAACAGGGAAGACUAUUUUAC |
| | GC/AA-mutant | CCCAGCCUGCCUGGGCUCUGACCACUAACACUCUUGGCCAUGGACAGCCUGCACAGGACCGCCUCCCUGCUCUUGGCCACUGUGCUC<br>CCAUUUCUGUCCUUGGCCUUGGGAGUAGCUGAGAUUUAAAAAAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUG<br>AUGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAGAGCAGACAGCAUCGCAUCGCACACUUCAUCACUGGCUGCCGGUCACAUC<br>CUUGGAGGGACUAAGCUGGGGGUGGGGGACAUGAGUCCCCCUGCUGCCCCUGCCACAUCCCAGUGGGCUCUGACCCCCUGAUCUCAA<br>CUCGUGGCACUAACUUGGAAAAGGGUUGAUUUAAACUAACAGGGAAGACUAUUUUAC |
| **LRP5 Ty PG4** | WT | CCUCGGCCGGGCCACUCUGGCUUCUCUGUGCCCCUGUAAAUAGUUUUAAAUAUGAACAAAGAAAAAAAUAUAUUUUAUGAUUUAAAA<br>AAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUGAUGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGGA<br>GAACUUUGUACAGUGGAGAAAUAUUUAUAAACUUAAUUUUGUAAAACAG |
| | G/A-mutant | CCUCGGCCGGGCCACUCUGGCUUCUCUGUGCCCCUGUAAAUAGUUUUAAAUAUGAACAAAGAAAAAAAUAUAUUUUAUGAUUUAAAA<br>AAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUGAUGGGAGCAGACUAGAGAGUAGAGCUGGCAGAGGAGAGAGA<br>GAACUUUGUACAGUGGAGAAAUAUUUAUAAACUUAAUUUUGUAAAACAG |

**Supplementary table S4** Oligodeoxynucleotides used to build full length 3'-UTR and directed mutagenesis

| **Constructions** | **Oligodeoxynucleotides Name** | **Sequence 5'-3'** |
|---|---|---|
| **TTYH1 WT and G/A-mutant** | TTYH1 3UTR Fow | ATCTCAGTCTAGACCCAGCCTGCCTGGGCTCTGACCACTAACACTCTTGGCCATGGACAGCCTGCAC<br>AGGACCGCCTCCC |
| | TTYH1 3UTR Rev | TACTCGAGGATCCGTAAAATAGTCTTCCCTTTTATTTTAAATCAACCCTTTTCCAAGTTAGTGCCAC<br>GAGTTG |
| | TTYH1 3UTR-1 | CCTGCACAGGACCGCCTCCCTGCTCTTGGCCACTGTGCTCCCATTTCTGTCCTTGGCCTTGGGAGTA<br>GCTGAGGG |
| | TTYH1 3UTR-2 Wt | GGCAGCCAGGGATGAAGGGTGCGAGGCGAGGCTGTCTGCCCCCTCCCCTGCCAGCCCTACTCCCTAG<br>TCTGCCCCCTCAGCTACTCCC |
| | TTYH1 3UTR-2 Mut | GGCAGCCAGGGATGAAGGGTGCGAGGCGAGGCTGTCTGCTCTCTCCTCTGCCAGCTCTACTCTCTAG<br>TCTGCTCCCTCAGCTACTCCC |
| | TTYH1 3UTR-3 | CCTCGCCTCGCACCCTTCATCCCTGGCTGCCGGTCCCATCCTTGGAGGGACTAAGCTGGGGGTGGGG<br>GACATGAGTCCCC |

| | | |
|---|---|---|
| | TTYH1 3UTR-4 | CCAAGTTAGTGCCACGAGTTGAGATCAGGGGGTCAGAGCCCACTGGGATGTGGCAGGGGCAGCAGGGGGACTCATGTCCCC |
| **TTYH1 + pAS** | TTYH1+pA Fow | CCTGCACAGGACCGCCTCCCTGCAATAAATCTTGGCCACTGTGCTCCC |
| | TTYH1 3UTR Rev-pAmut | TACTCGAGGATCCGTAAAATAGTCTTCCCTGTTAGTTTAAATCAACCCTTTTCCAAGTTAGTGCCACGAGTTG |
| **LRP5 Ty PG4 WT and G/A-mutant** | TyGquad Wt-lrp5 Fow | GTGAACTGTGATGGGGGCAGACTAGGGAGTAGGGCTGGCAGGGGAGGGGGAGAACTTTGTACAGTGG |
| | TyGquad Wt-lrp5 Rev | CCACTGTACAAAGTTCTCCCCCTCCCCTGCCAGCCCTACTCCCTAGTCTGCCCCCATCACAGTTCACATTTCTC |
| | TyGquad Mut-lrp5 fow | GTGAACTGTGATGGGAGCAGACTAGAGAGTAGAGCTGGCAGAGGAGAGAGAGAACTTTGTACAGTGG |
| | TyGquad Mut-lrp5 Rev | CCACTGTACAAAGTTCTCTCTCTCCTCTGCCAGCTCTACTCTCTAGTCTGCTCCCATCACAGTTCACATTTCTC |
| **TTYH1-LRP5-pAS WT and G/A-mutant** | LRP5-partUTR Ty_Wt Fow | GGGATTTTAAAAACATGAGAAATGTGAACTGTGATGGGGGCAGACTAGGGAGTAGGGCTGGCAGGGGAGGGGGCAGACAGC |
| | LRP5-partUTR Ty_Mut Fow | GGGATTTTAAAAACATGAGAAATGTGAACTGTGATGGGAGCAGACTAGAGAGTAGAGCTGGCAGAGGAGAGAGCAGACAGC |
| | LRP5-partUTR Ty Rev | CACATTTCTCATGTTTTTAAAATCCCAATTATATTTATTTTTTAAATCTCAGCTACTCCCAAGGCCAAGG |
| **TTYH1 C/A-mutant and GC/AA-mutant** | Ty-C Fow | CAGCATCGCATCGCACACTTCATCACTGGCTGCCGGTCACATCCTTGGAGGGACTAAGCTGG |
| | Ty-C Wt Rev | GGATGTGACCGGCAGCCAGTGATGAAGTGTGCGATGCGATGCTGTCTGCCCCCTCCCCTGCCAGCCC |
| | Ty-C Mut Rev | GGATGTGACCGGCAGCCAGTGATGAAGTGTGCGATGCGATGCTGTCTGCTCTCTCCTCTGCCAGCTC |
| **LRP5 WT and G/A-mutant** | LRP5-1 | TCAGTCTAGACCTCGGCCGGGCCACTCTGGCTTCTCTGTGCCCCTGTAAATAGTTTTAAATATGAACAAAGAAAAAAATATATTTTATGATTTAAAAAAT |
| | LRP5-2Wt | GTACAAAGTTCTCCCAGCCCTGCCCACCCCATCACAGTTCACATTTCTCATGTTTTTAAAATCCCAATTATATTTATTTTTTAAATCATAAAATATATTT |
| | LRP5-3 | TCGAGGATCCCTGTTTTACAAAATTAAGTTTATAAATATTTCTCCACTGTACAAAGTTCTC |
| | LRP5-2Mut | GTACAAAGTTCTCTCAGCTCTGCTCACTCCATCACAGTTCACATTTCTCATGTTTTTAAAATCCCAATTATATTTATTTTTTAAATCATAAAATATATTT |
| | LRP5 3UTR Rev XbaI | TCGATCTAGACTGTTTTACAAAATTAAGTTTATAAATATTTCTCCAC |
| | LRP5-1 Fow* | TCAGTCTAGACCTCGGCCGGGCCACTCTGGCTTCTCTGTGCCCCTGTAAATAGTTTTAAATATGAACAAAG |

**Supplementary table S5** Area under the curve of the different predictive parameters

| | Total loop length | Mfe | cG/cC | QGRS G-score |
|---|---|---|---|---|
| **Area under the curve (AUC)** | 0,6795 | 0,7564 | 0,9679 | 0,8974 |

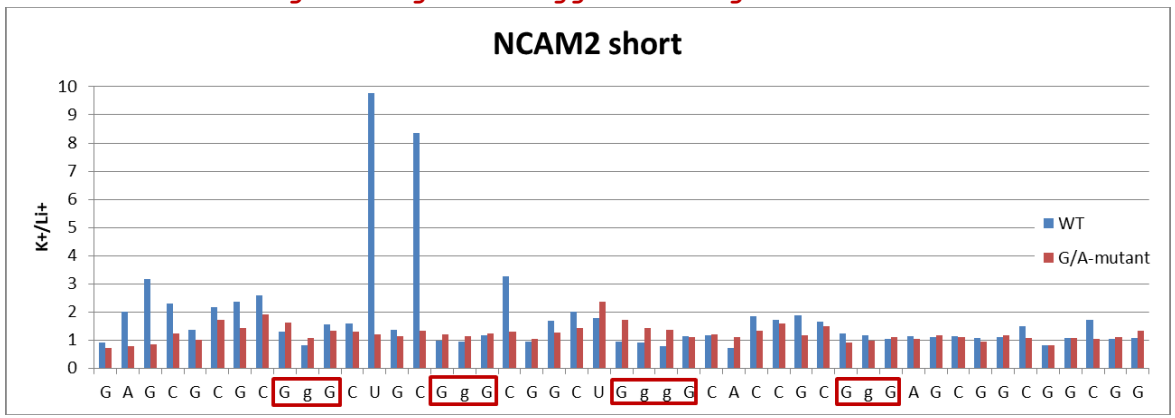
**Supplementary table S6** Table of sensitivity and specificity percentages for the different cG/cC score thresholds

| Threshold | Sensitivity (%) | Specificity (%) | False positives (FP) | False negatives (FN) | True positives (TP) | True negatives (TN) |
|---|---|---|---|---|---|---|
| > 0.70 | 100 | 16,67 | 7 | 0 | 6 | 1 |
| > 1.00 | 100 | 50 | 6 | 0 | 6 | 2 |
| > 1.60 | 100 | 66,67 | 5 | 0 | 6 | 3 |
| > 2.05 | 100 | 83,33 | 3 | 0 | 6 | 5 |
| > 2.15 | 92,31 | 83,33 | 3 | 0 | 6 | 5 |
| > 2.40 | 84,62 | 83,33 | 3 | 1 | 5 | 5 |
| > 2.70 | 76,92 | 100 | 2 | 1 | 5 | 6 |
| > 3.05 | 69,23 | 100 | 1 | 1 | 5 | 7 |
| > 3.35 | 61,54 | 100 | 0 | 2 | 4 | 8 |
| > 3.70 | 53,85 | 100 | 0 | 2 | 4 | 8 |
| > 4.25 | 46,15 | 100 | 0 | 2 | 4 | 8 |
| > 4.70 | 38,46 | 100 | 0 | 2 | 4 | 8 |
| > 5.10 | 30,77 | 100 | 0 | 2 | 4 | 8 |
| > 6.00 | 23,08 | 100 | 0 | 3 | 3 | 8 |
| > 7.15 | 15,38 | 100 | 0 | 4 | 2 | 8 |
| > 9.20 | 7,69 | 100 | 0 | 4 | 2 | 8 |

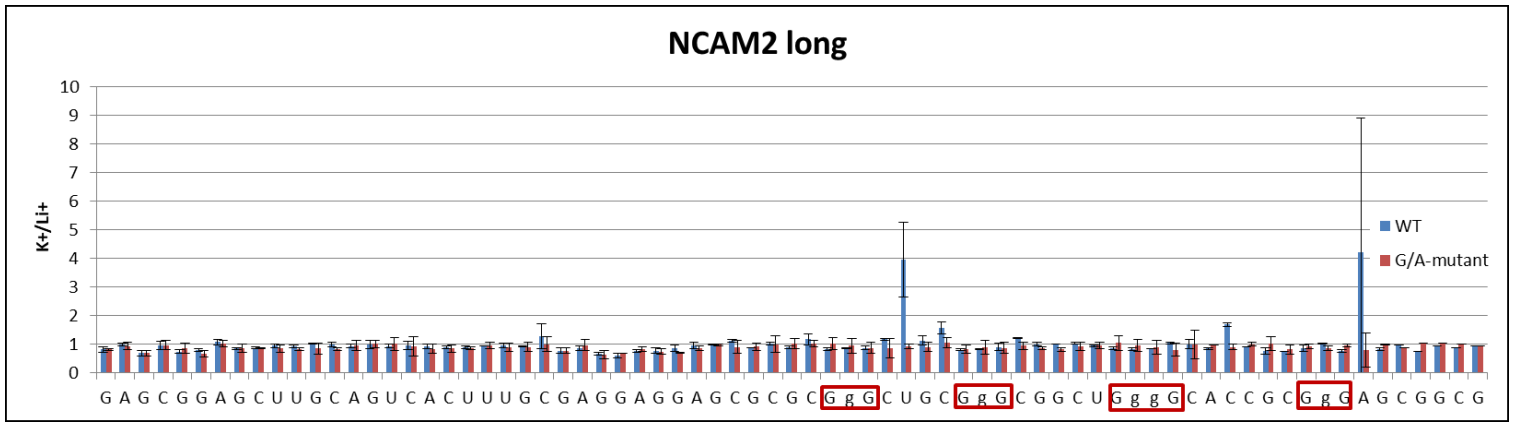
**Supplementary Figures S1 to S27**

**S1**

5'-GGAGGAGCGCGC**GgGCUGCGgGCGGCUGggGCACCGCGgG**AGCGGCGGCGGCGG-3'

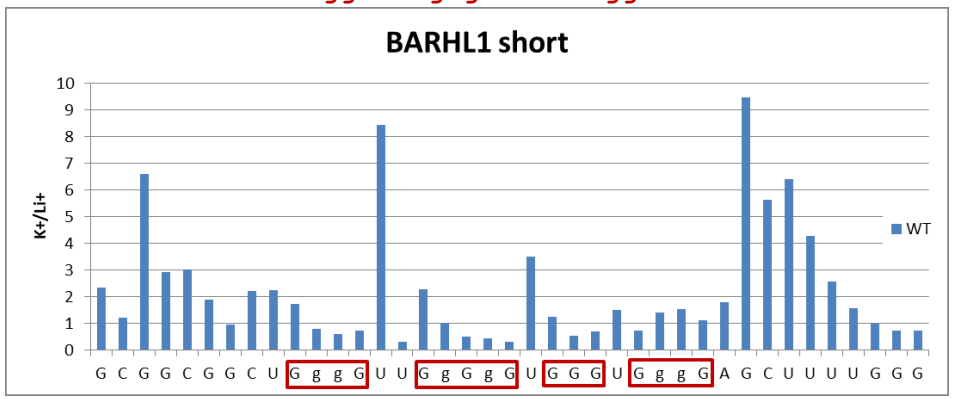



5'-
GGGAUAGUGCGGCAAGAGCGGAGCUUGCAGUCACUUUGCGAGGAGGAGCGCGC**GgGCUGCGgGCGGCUGggGCACCGCGgG**AGCGGCGGCGGCGGCUC
UAGCAGAGGCGGCCGGGGCAGCGAAAGGUUCUC-3'

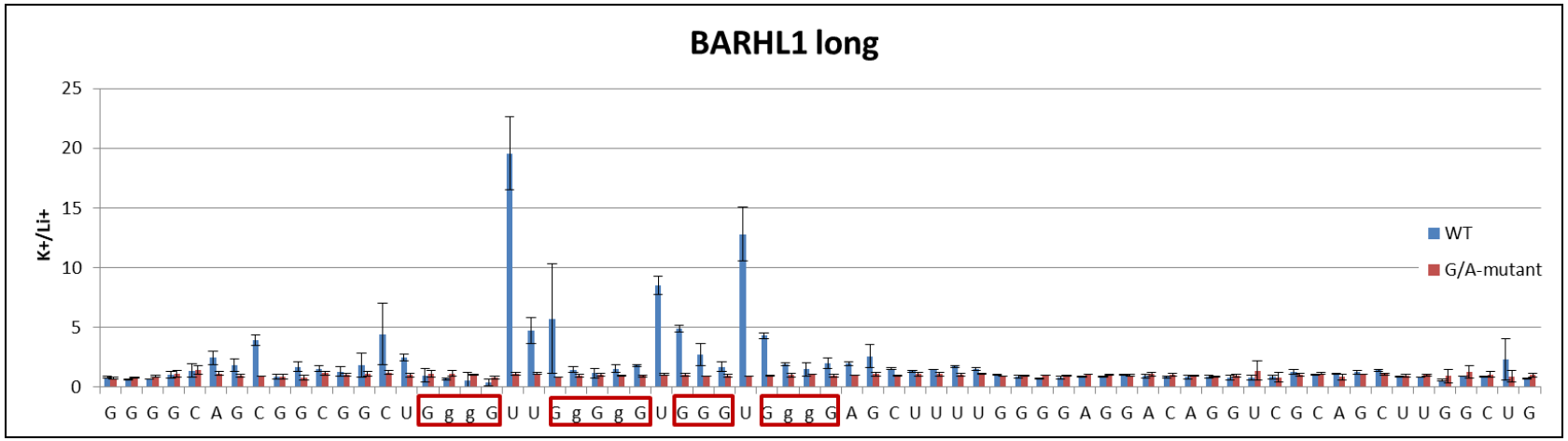


**S2**

5'GGGGCAGCGGCGGCU**GggGUUGgGgGUGGGUGggG**AGCUUUUGGGG-3'

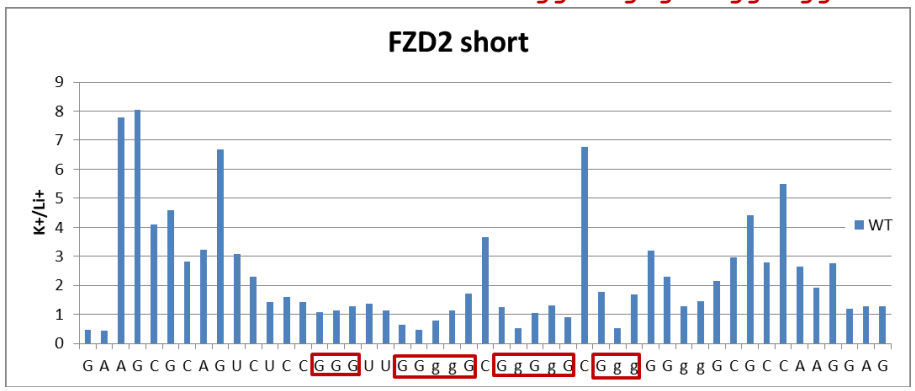



5'-
GGGCCCGCCUUCCCCAUGCCAGCCCGCAGCUAGGGGCAGGGGCAGCGGCGGCU**GggGUUGgGgGUGGGUGggG**AGCUUUUGGGGAGGACAGGUCGCAG
CUUGGCUGCUAGCCACCAUGACUUC-3'

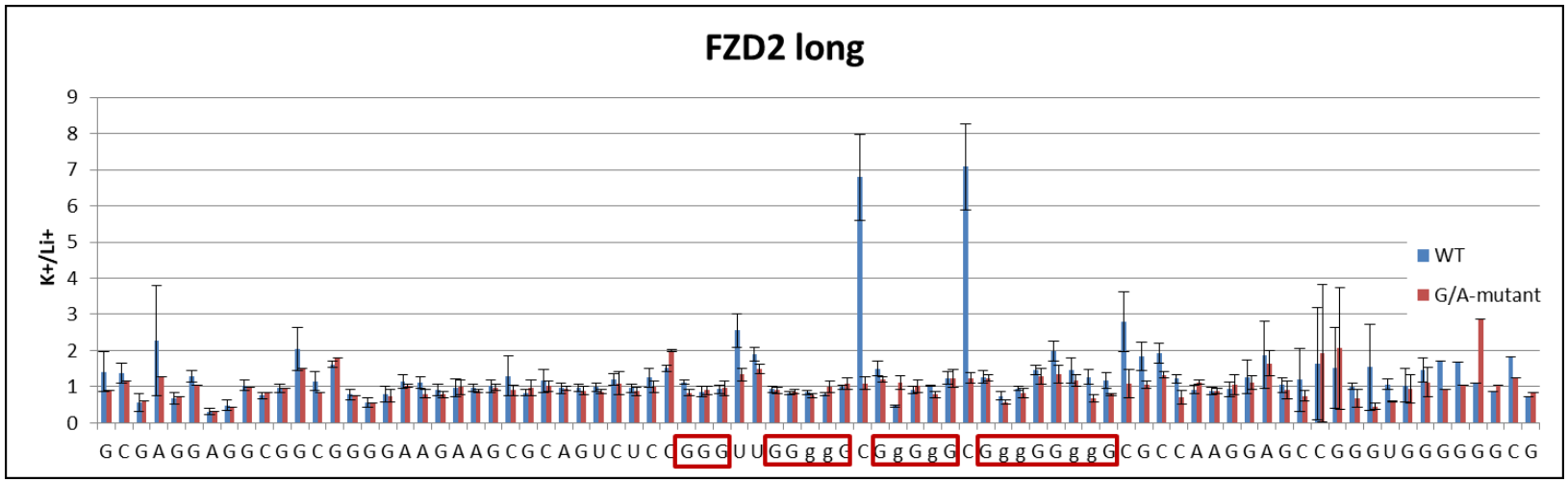


**S3**

5'-GGGGAAGAAGCGCAGUCUCC**GGGUUGGggGCGgGgGCGggGGggG**CGCCAAGGAGCCGGG-3'

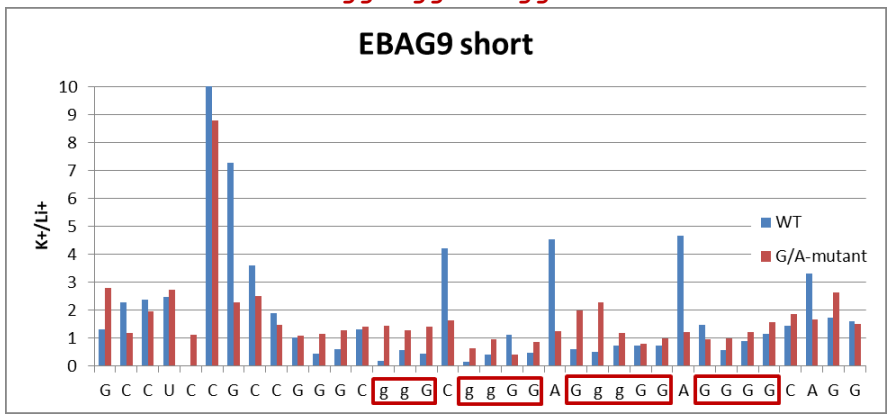



5'-
GGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGUCUCC**GGGUUGGggGCGgGgGCGggGGggG**CGCCAAGGAGCCGGGUGGGGGGGCG
GCGGCCAGCGCUAGCCACCAUGACU-3'

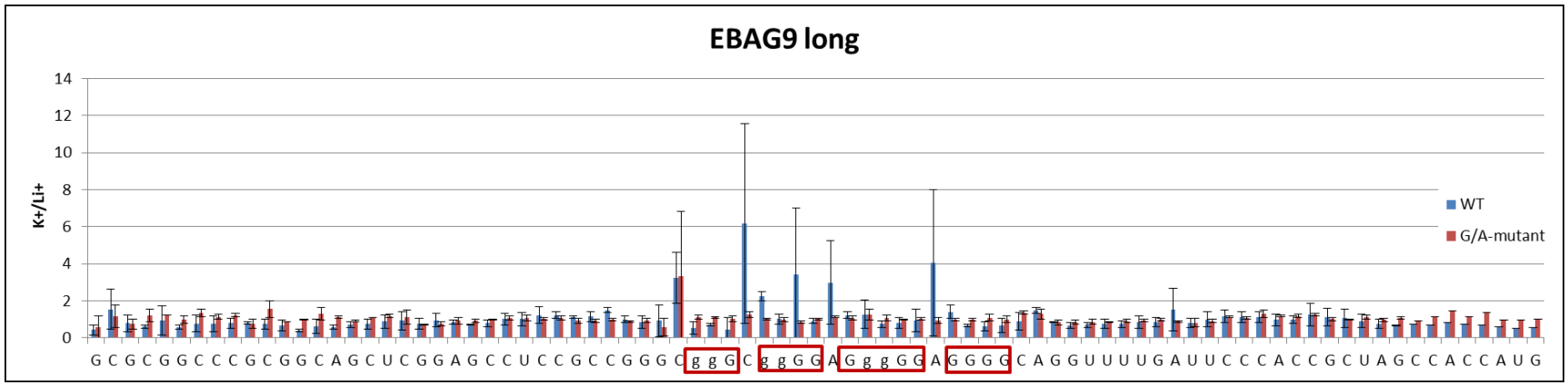


**S4**

5'-GGAGCCUCCGCC**GGGC**gg**GC**gg**GGAG**gg**GGAGGGG**CAGGUUUUGA-3'

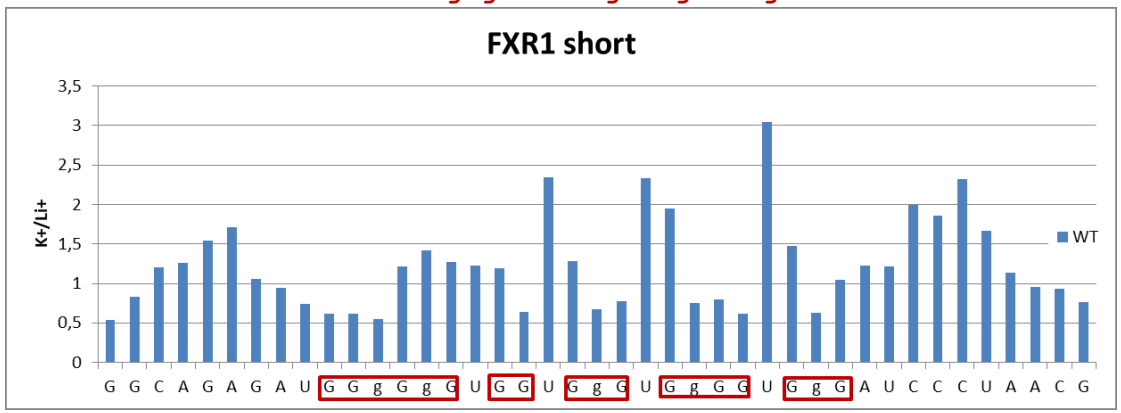



5'-
GGGAGCGCGCCUUGUGUGCGCGCGCGGCCCGCGGCAGCUCGGAGCCUCCGCC**GGGC**gg**GC**gg**GGAG**gg**GGAGGGG**CAGGUUUUGAUUCCCACCGCUAG
CCACCAUGACUUCGAAAGUUUAUGAUC3'

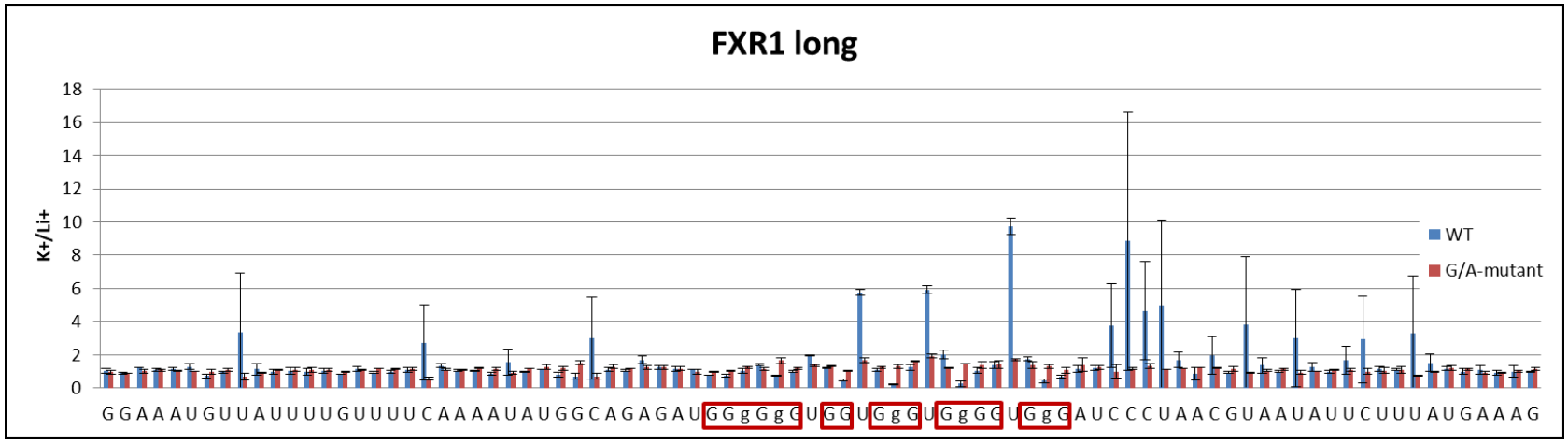


**S5**

5'-GGGCAAAAUAUGGCAGAGAUGGgGgGUGGUGgGUGgGGUGgGAUCCCUAACGUAAUA-3'

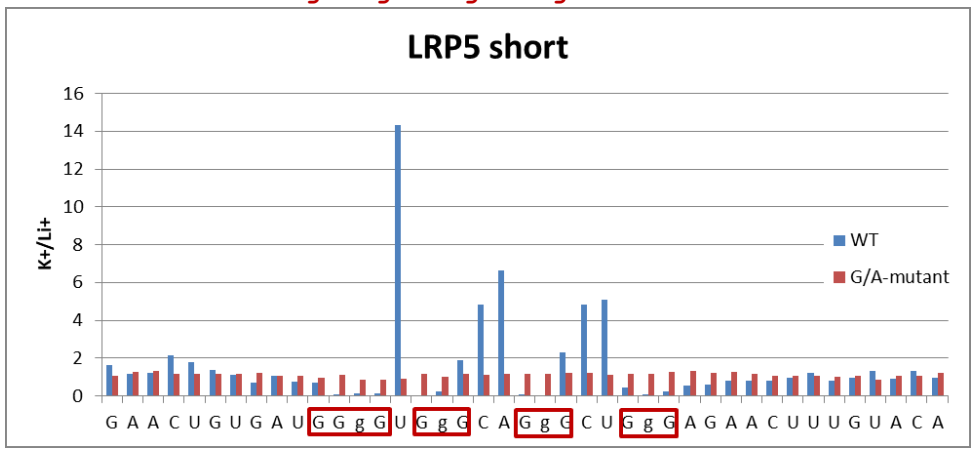



5'-
GGGUUGCUGGCUAUAGGAAAUGUUAUUUUGUUUUCAAAAUAUGGCAGAGAUGGgGgGUGGUGgGUGgGGUGgGAUCCCUAACGUAAUAUUCUUUAUGA
AAGCAUUAGCUGCUUUUGUUACAUU-3'

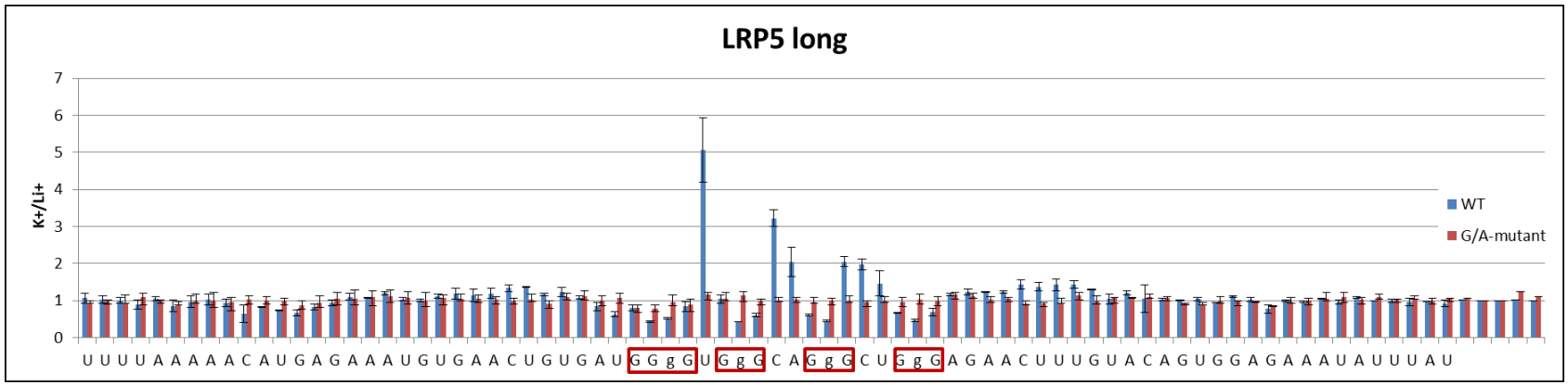


**S6**

5'-GGGAACUGUGAUGGgGUGgGCAGgGCUGgGAGAACUUUGUACAG-3'

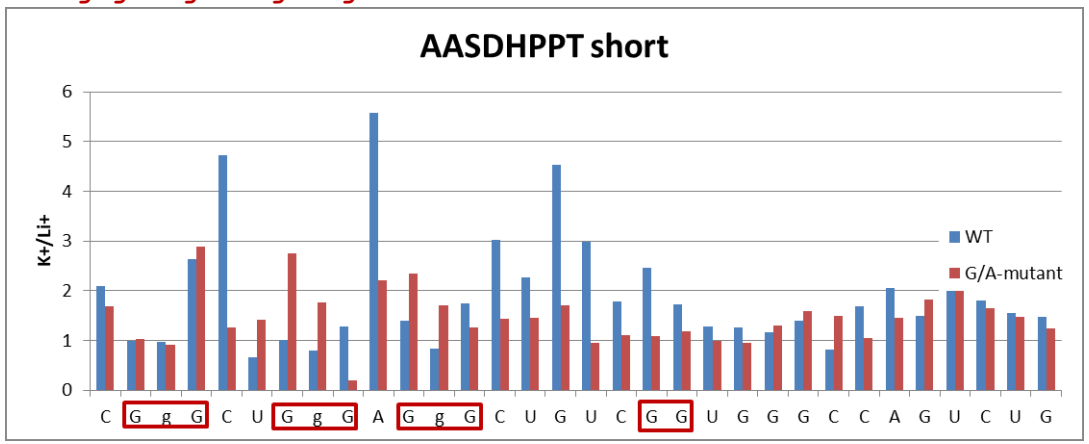



5'-
GGGAAAAUAAAUAUAAUUGGGAUUUUAAAAACAUGAGAAAUGUGAACUGUGAUGGgGUGgGCAGgGCUGgGAGAACUUUGUACAGUGGAGAAAUAUUU
AUAAACUUAAUUUUGUAAAACAG-3'

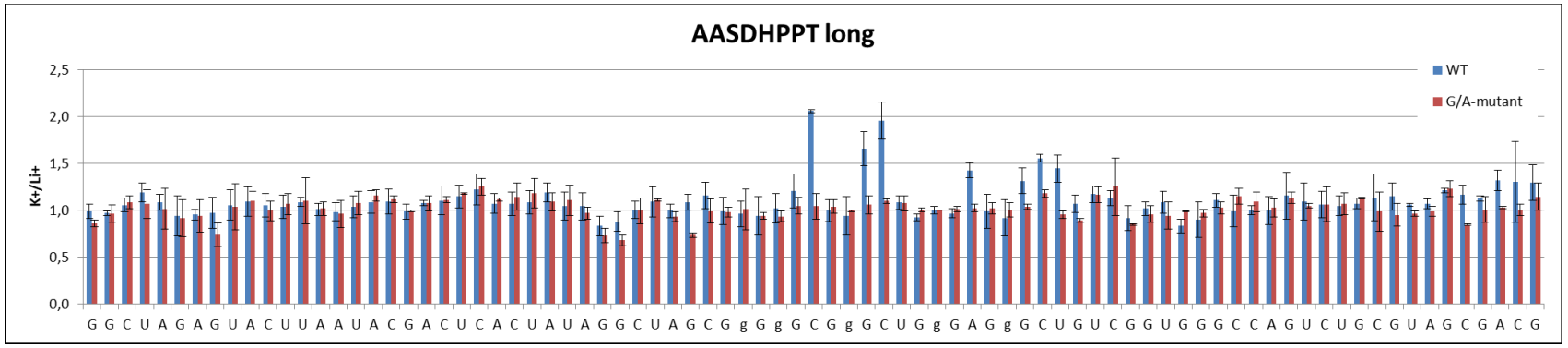


**S7**

5'-GgGgGCGgGCUGgGAGgGCUGUCGGUGGGCCAGUCUGC-3'

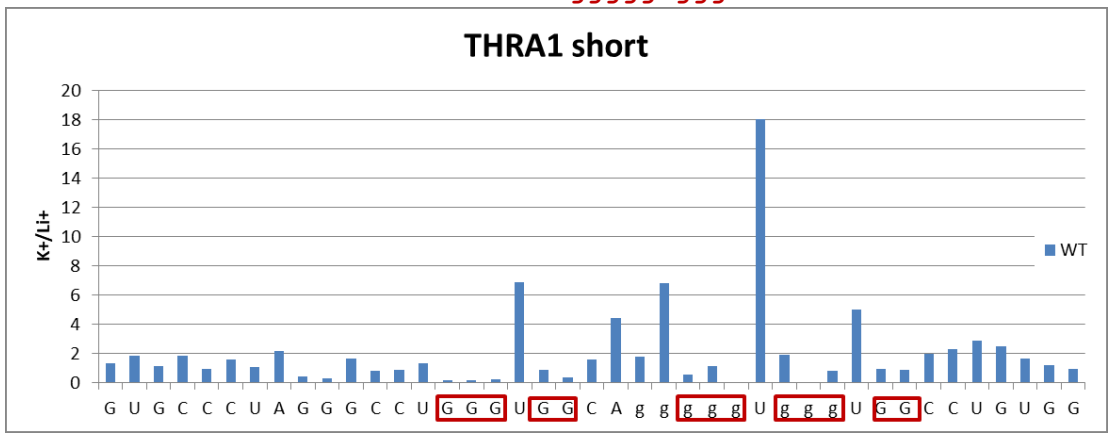



5'-
GGGAUUACAGCUCUUAAGGCUAGAGUACUUAAUACGACUCACUAUAGGCUAGCGgGgGCGgGCUGgGAGgGCUGUCGGUGGGCCAGUCUGCGUAGCGA
CGGCCCGUCCCCUGCGCACGGAC-3'

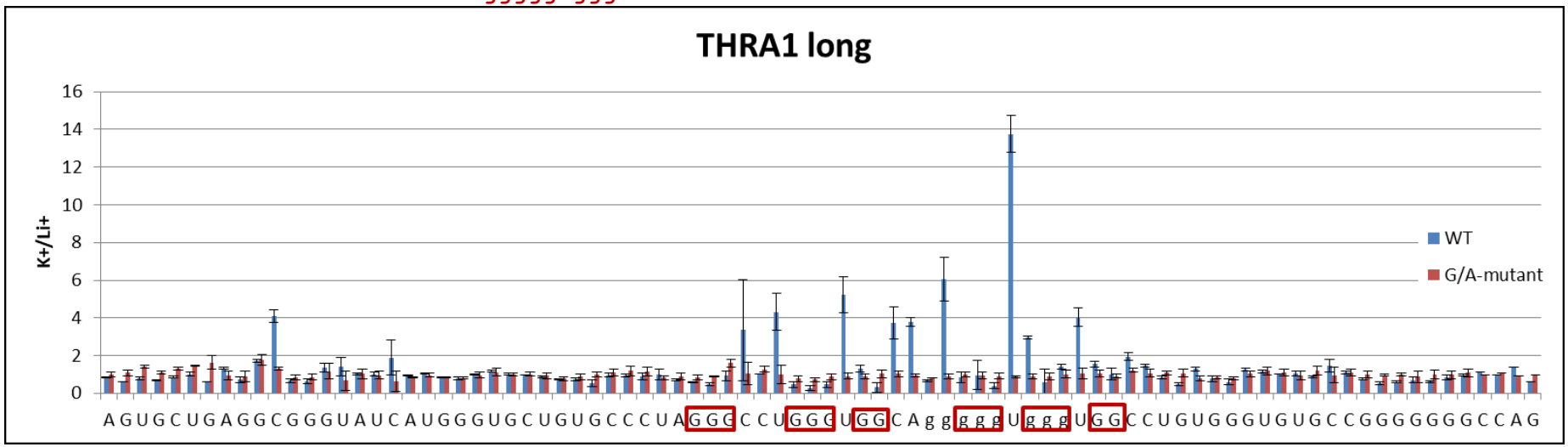


**S8**

5'-GGGUGCUGUGCCCUA**GGGCCUGGGUGGCA**g**ggggg**U**ggg**UGGCCUGUGGG-3'

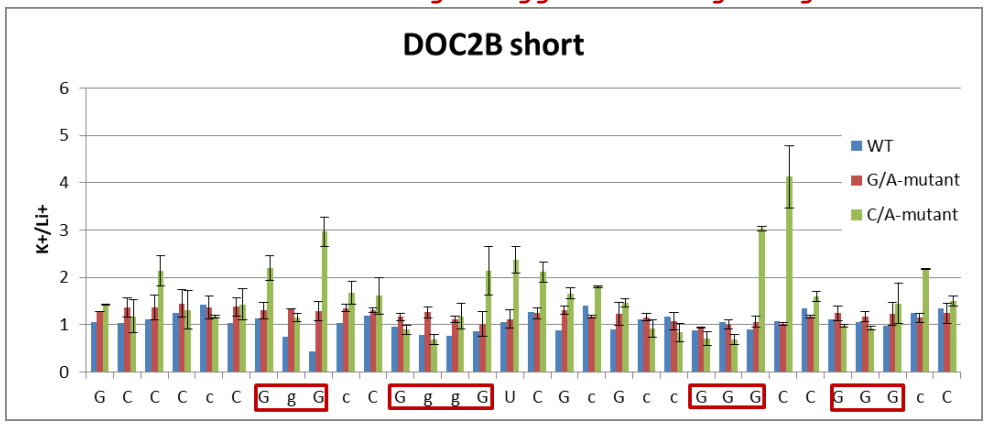



5'-GGGUGCUGUGCCCUA**GGGCCUGGGUGGCA**g**ggggg**U**ggg**UGGCCUGUGGG-3'

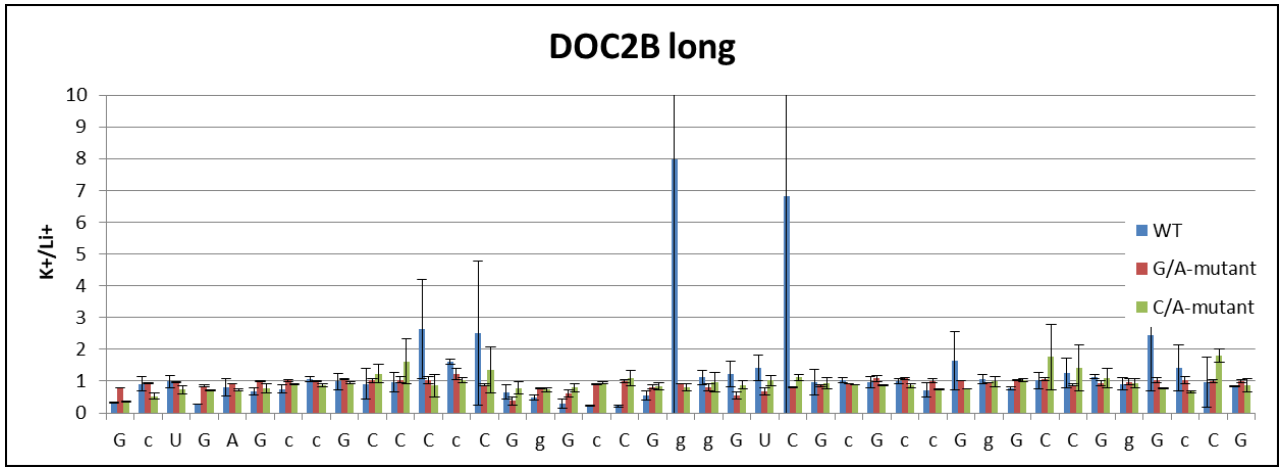


**S9**

5'-GGGccUGcUGAGccGCCCcCGgGcCGggGUCGcGccGgGCCGgGcCGCGCcCGGGG-3'

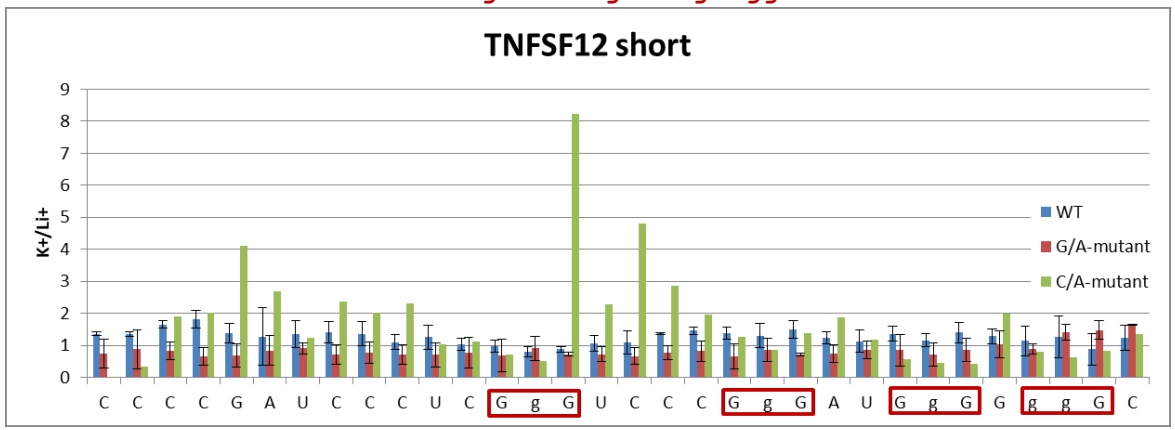



5'-
GGGCGAUGCCCGCAGCCCCCGCCGCGCCCCGCCGGGccUGcUGAGccGCCCcCGgGcCGggGUCGcGccGgGCCGgGcCGCGCcCGGGGCGGGGCGGC
GCUGCCUGCGCUAGCCACCAUGACUUCGA-3'

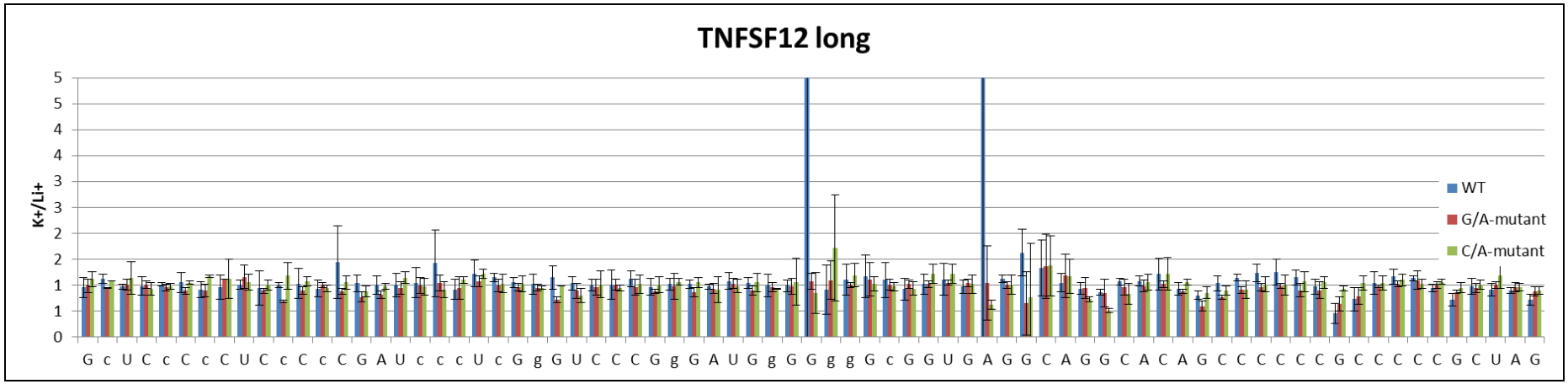


**S10**

5'-GGcUCcCcCUCcCcCGAUcccUcGgGUCCCGgGAUGgGGggGcGGUGAGGCAGG-3'

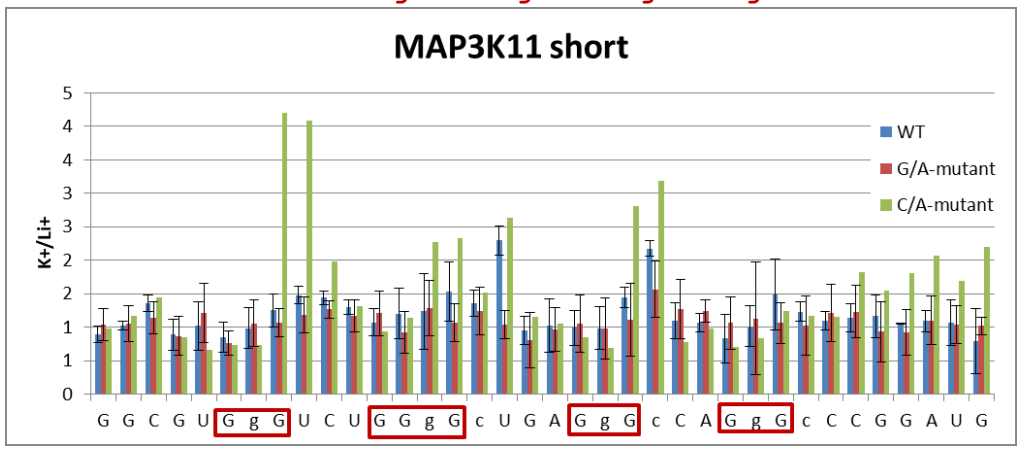



5'-
GGGCCUCUCCCCGGCCCGAUCCGCCCGCCGGcUCcCcCUCcCcCGAUcccUcGgGUCCCGgGAUGgGGggGcGGUGAGGCAGGCACAGCCCCCCGCCC
CCGCUAGCCACCAUGACUUCGAA-3'

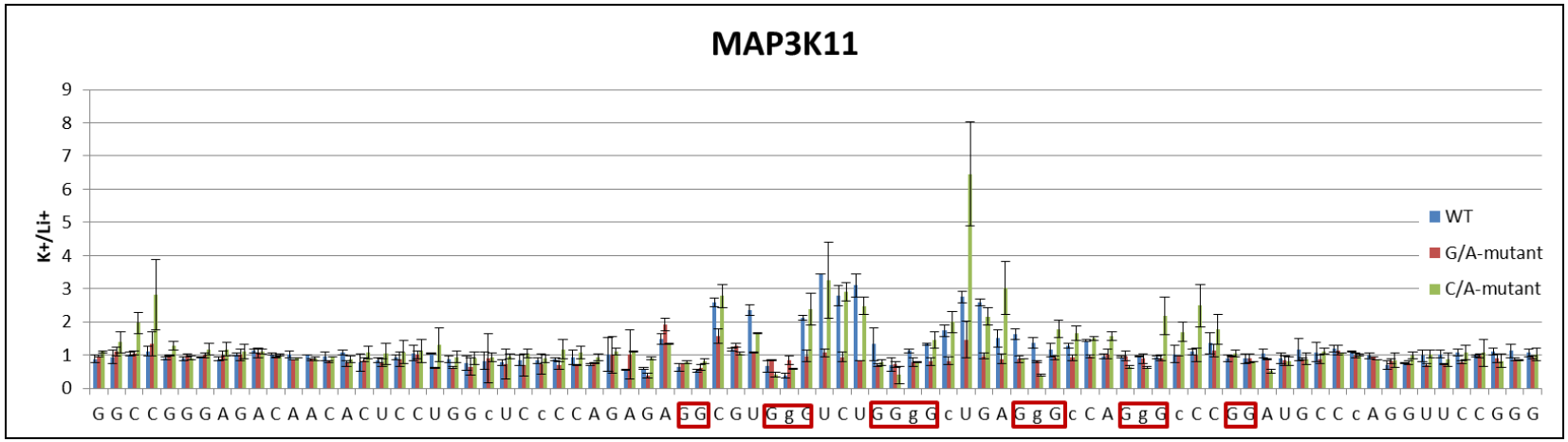


**S11**

5'-GGcUCcCCAGAGAGGCGUGgGUCUGGgGcUGAGgGcCAGgGcCCGGAUGCCcAGG-3'

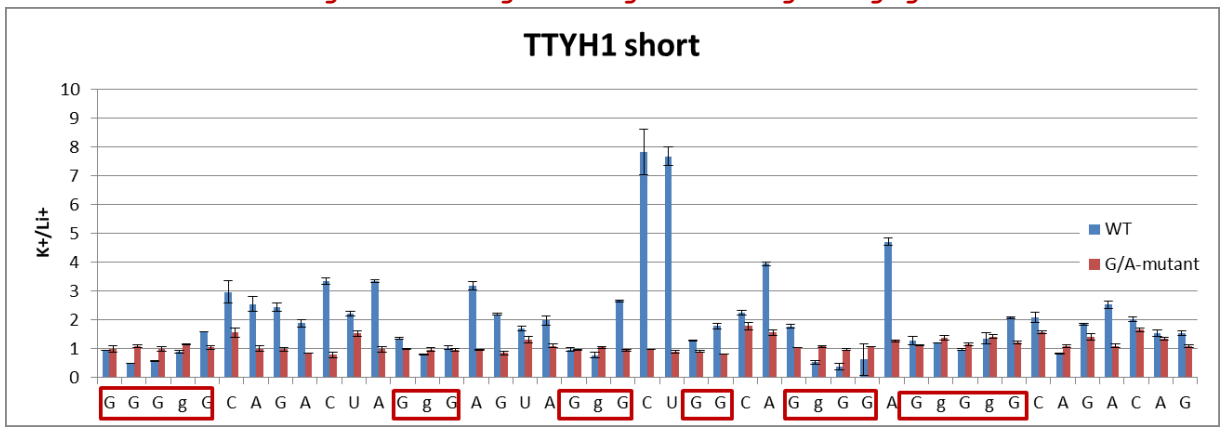



5'-
GGGCGAGAUGCGGGGGGCCGGGAGACAACACUCCUGGcUCcCCAGAGAGGCGUGgGUCUGGgGcUGAGgGcCAGgGcCCGGAUGCCCAGGUUCCGGGA
CUAGGGCCUUGGCAGCCAGCGGGGGUGG-3'

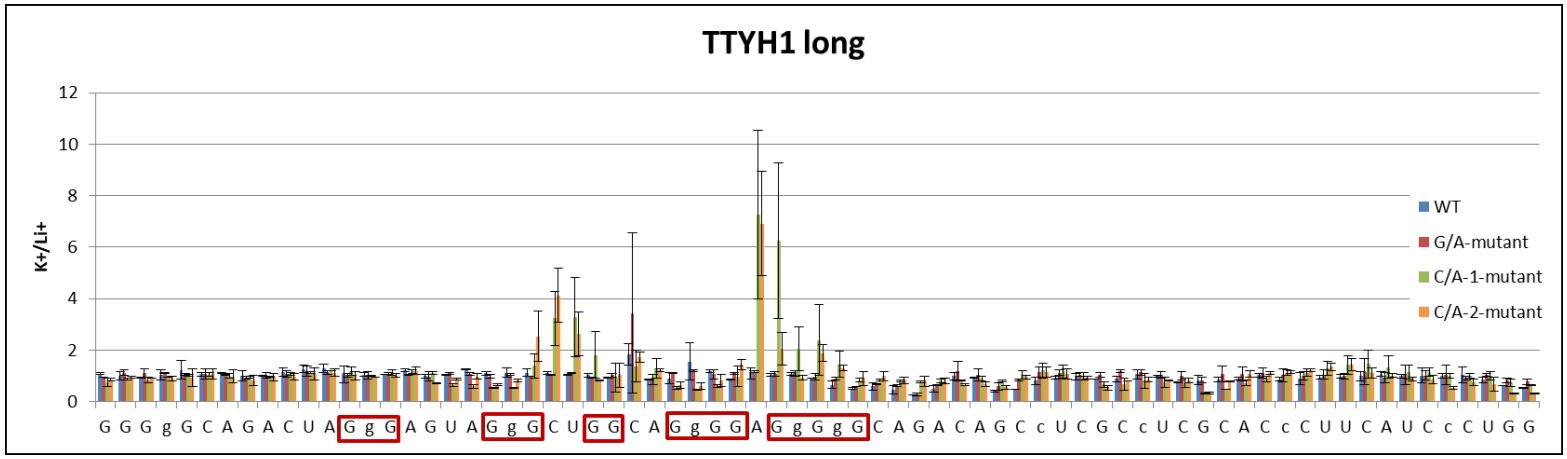


**S12**

5'-GGGAGUAGCUGAGGGgGCAGACUAGgGAGUAGgGCUGGCAGgGGAGgGgGCAGACAGCCUCGC-3'

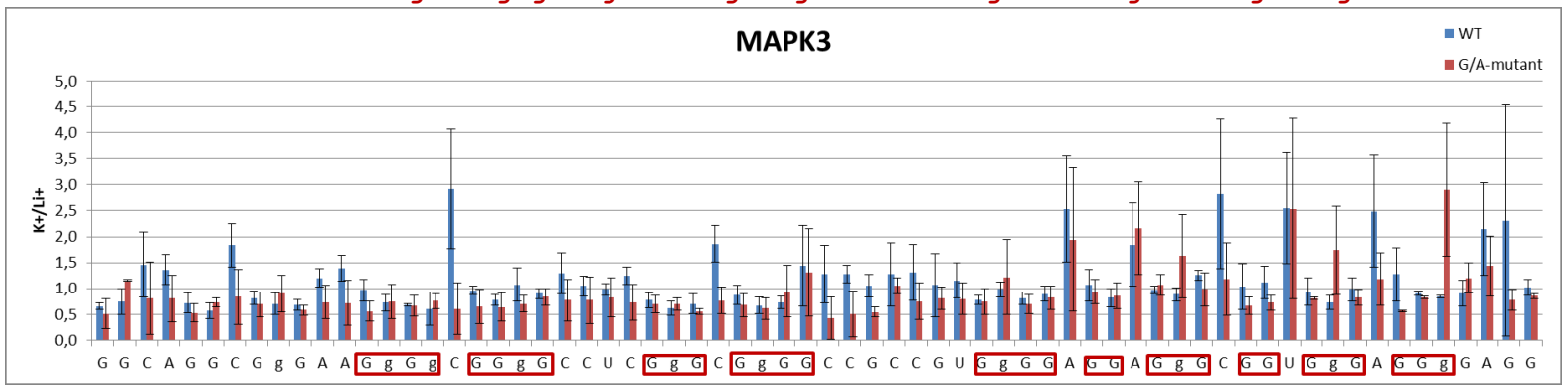



5'-
GGGUGcUCcCAUUUCUGUcCUUGGccUUGGGAGUAGCUGAGGGgGCAGACUAGgGAGUAGgGCUGGCAGgGGAGgGgGCAGACAGCcUCGCcUCGCAC
cCUUCAUCcCUGGCUGCCGGUCcCAUCCUU-3'



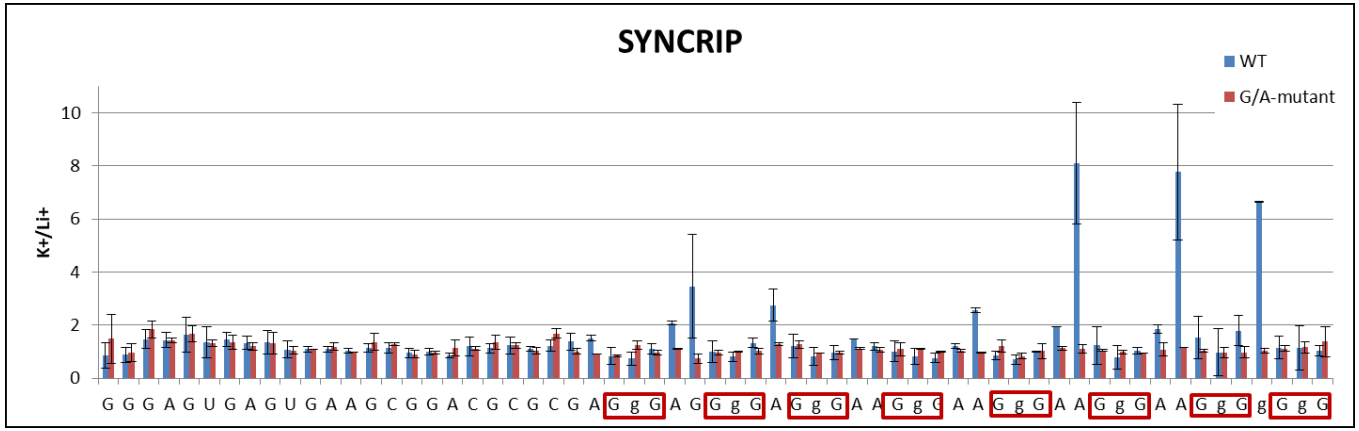
**S13**

5'-GGGCGGGUGACAGGCAGGCGgGAAGgGgCGGgGCCUCGgGCGgGGCCGCCGUGgGGAGGAGgGCGGUGgGAGGgGAGGAGUGGAG-3'



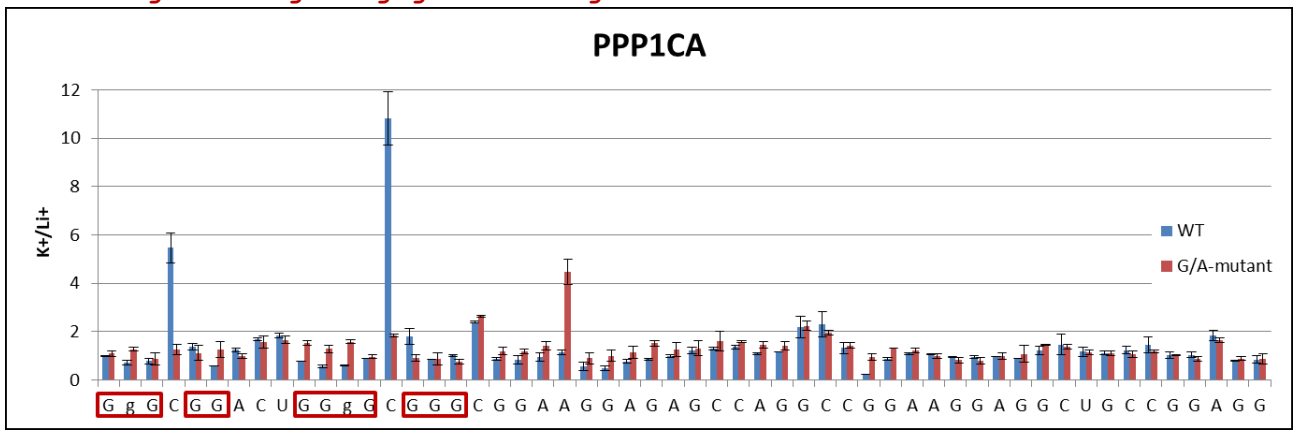

**S14**

5'-
GGGAGCUGGAGGAGGGCAGGGGCUGAGGGAGUGAGUGAAGCGGACGCGCGAGgGAGGgGAGgGAAGgGAAGgGAAGgGAAGgGgGgGUCACGCGgGgG
CGCGCGCGCGCACCGgGAGCGCGCUCGGAG-3'



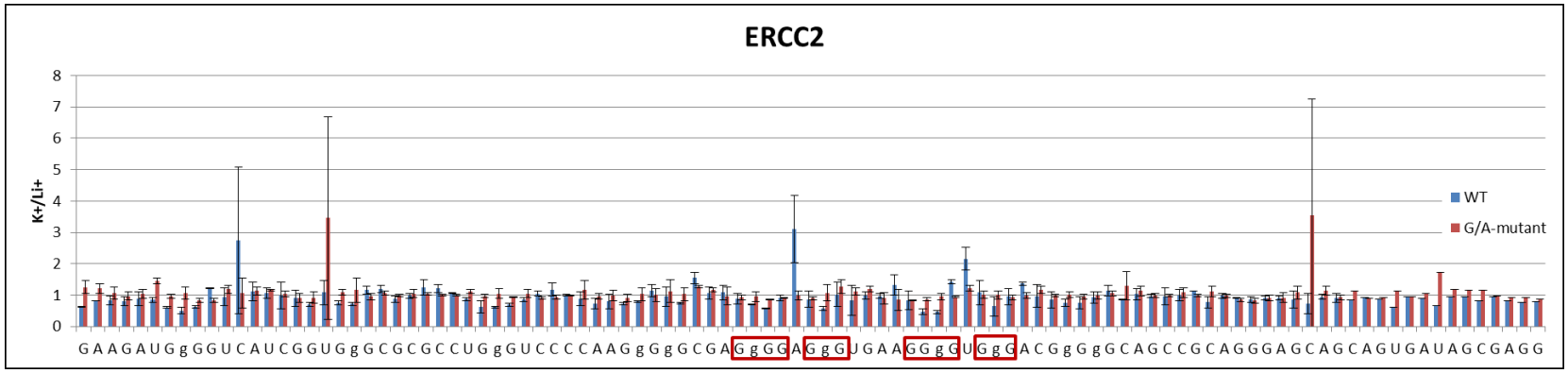
**S15**

5'-GGGCGgGGCCGCGgGCCGgGgGCGGACUGGgGCGGGCGGAAGGAGAGCCAGGCCGGAAGGAGGCUGCCGGAGGGCGGGAGG-3'



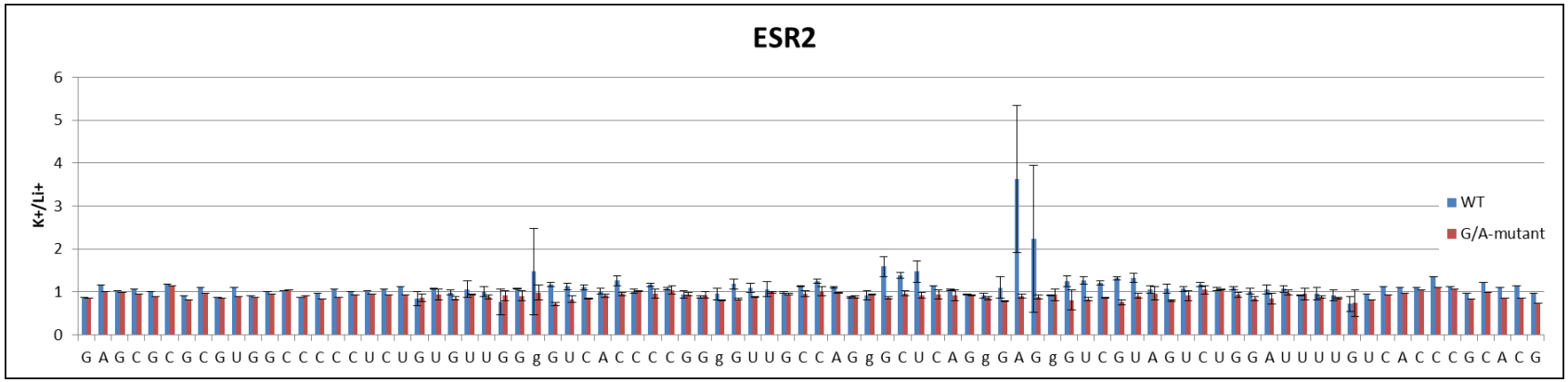

**S16**

5'-
GGGCGGgGGgUCUUGAAGAUGgGGUCAUCGGUGgGCGCGCCUGgGUCCCCAAGgGgGCGAGgGGAGgGUGAAGGgGUGgGACGgGgGCAGCCGCAGGGAGCAGCAGU
GAUAGCGAGGAGACACUGA-3'



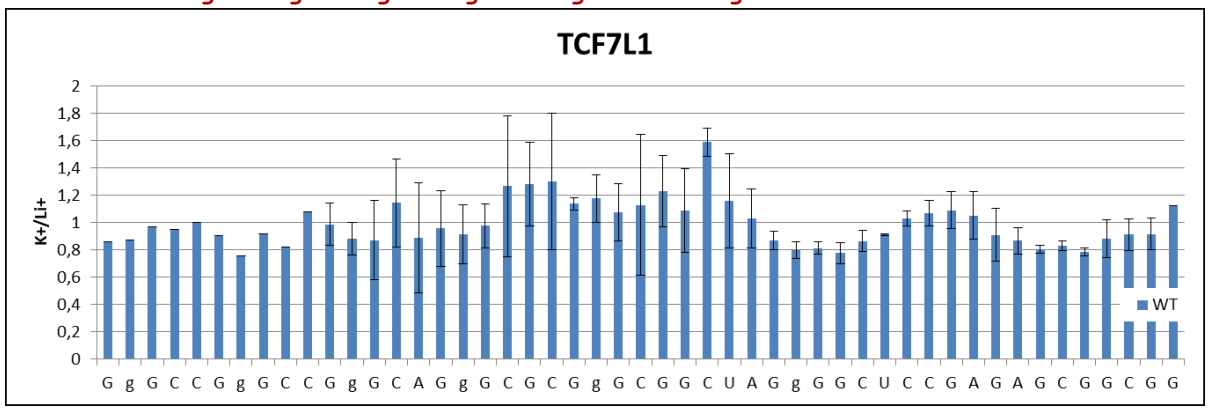
**S17**

5'-
GGGAGUGUCAGAGCUGGAGCGCGCGUGGCCCCCUCUGUGUU**GGgGUCACCCCGGgGUUGCCAGgGCUCAGgGAGgG**UCGUAGUCUGGAUUUUCUCACC
CGCACGUCCCCACCCCCCAGCAGGUCUG-3'



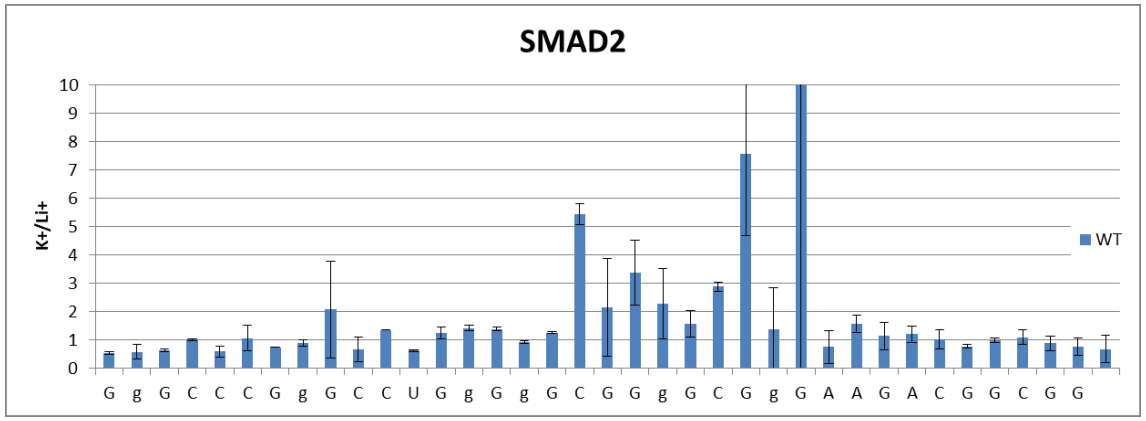

**S18**

5'-GGGCGCC**GgGCCGgGCCGgGCAGgGCGCGgGCGGCUAGgGG**CUCCGAGAGCGGCGGCCCCGGCCCGCGGCCCACC-3'



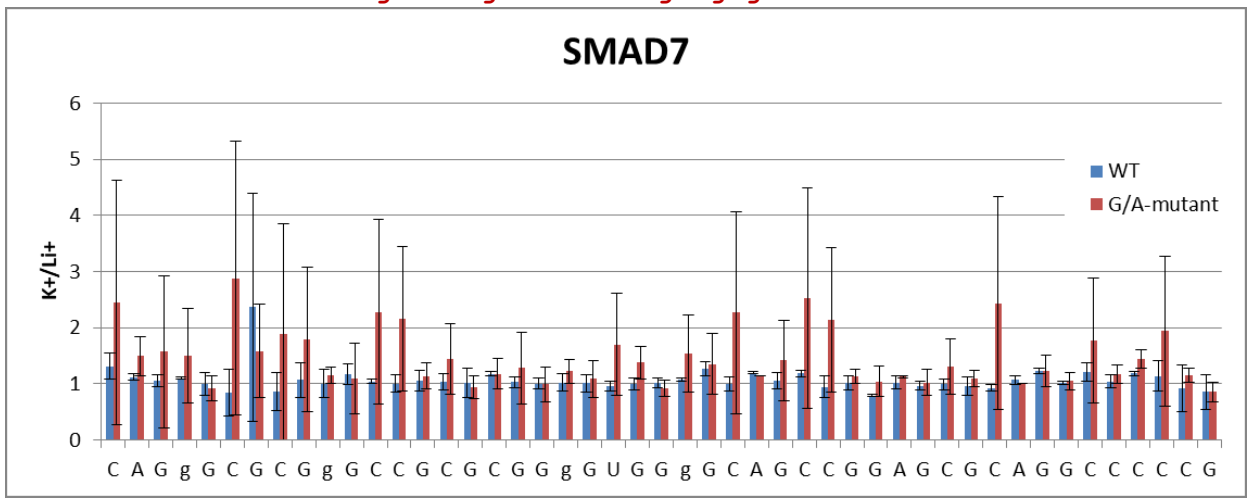
**S19**

5'-
GGGCGCCCGgGCCGCCGGCCGgGCCCGgGCCUGgGgGCGGgGCGgGAAGACGGCGGCCGGGAGUGUUUUCAGUUCCGCCUCCAAUCGCCCAUUCCC-
3'



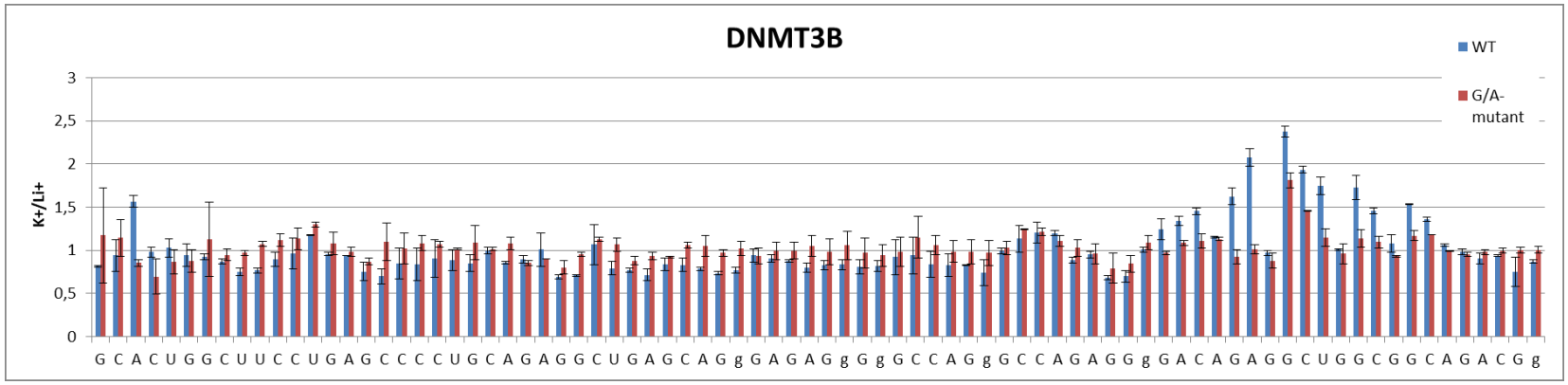

**S20**

5'-GGGCGGAGAGCCGCGCAGgGCGCGgGCCGCGCGGgGUgGgGCAGCCGGAGCGCAGGCCCCCGAUCCCCGGCGGGCGCCCCCGGGCCCCCGC-3'



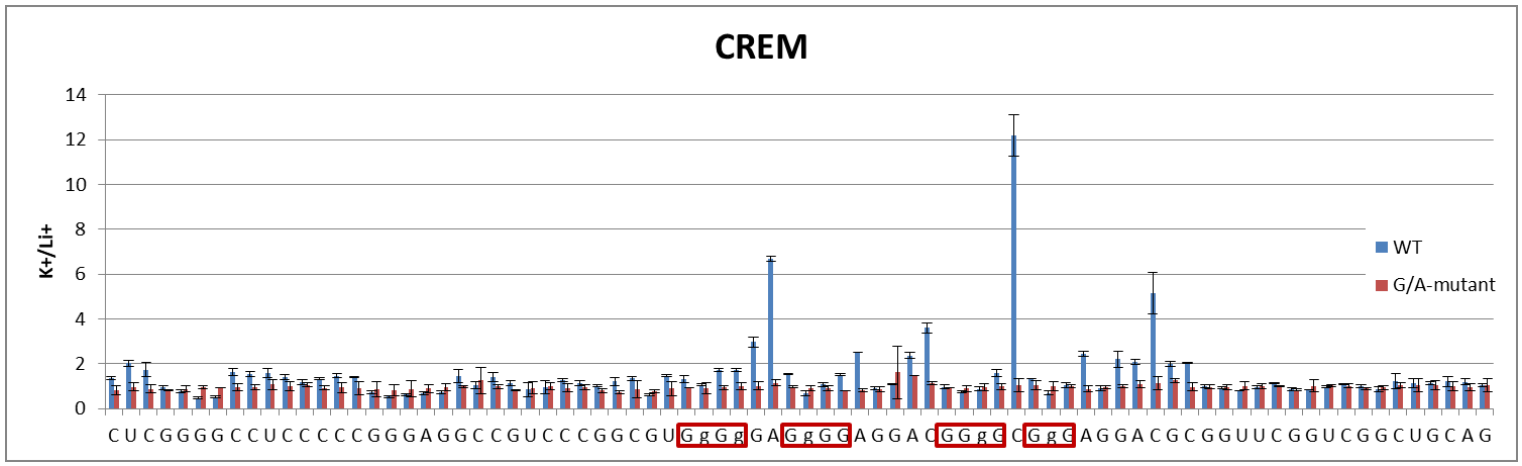
**S21**

5'-
GGGAGGAGGGAAAAUAAUGCACUGGCUUCCUGAGCCCCUGCAGAGGCUGAGCAGgGAGAGgGgGCCAGgGCCAGAGGgGACAGAGGCUGGCGGCAGACGgGCCGGGA
CAGGCAGGUCCUAAAUGGCAUU-3'



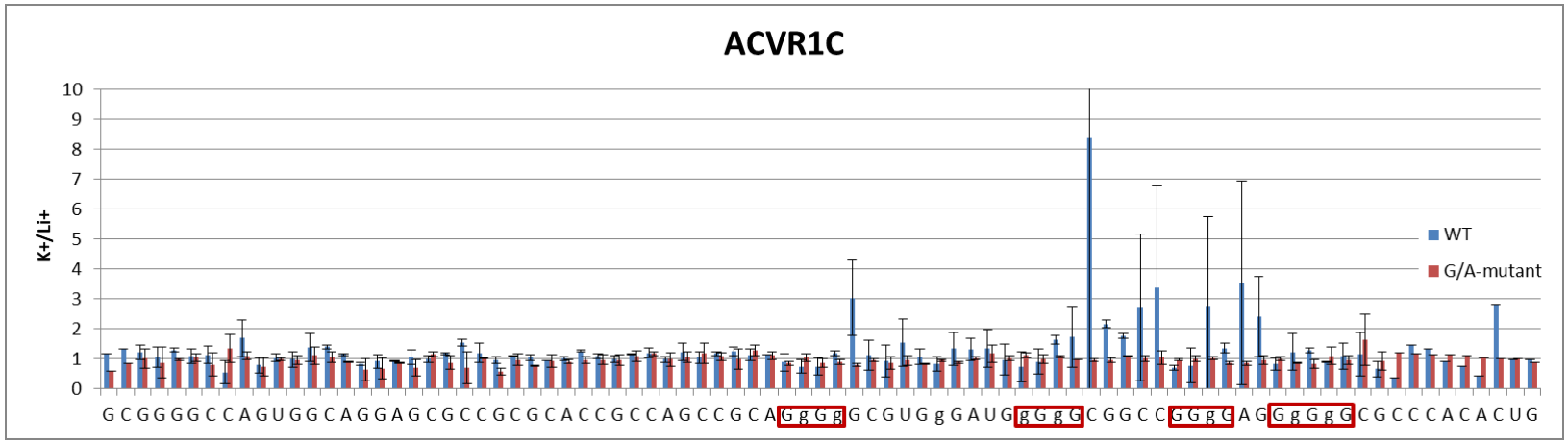

**S22**

5'-
GGGAGCCUGGAUUUUUUUCCUCGGGGCCUCCCCCGGGAGGCCGUCCCGGCGUGgGgGAGgGGAGGACGGgGCGgGAGGACGCGGUUCGGUCGGCUGCAGCGCUACUU
UUGGUCCGGGGUCGGCAG-3'



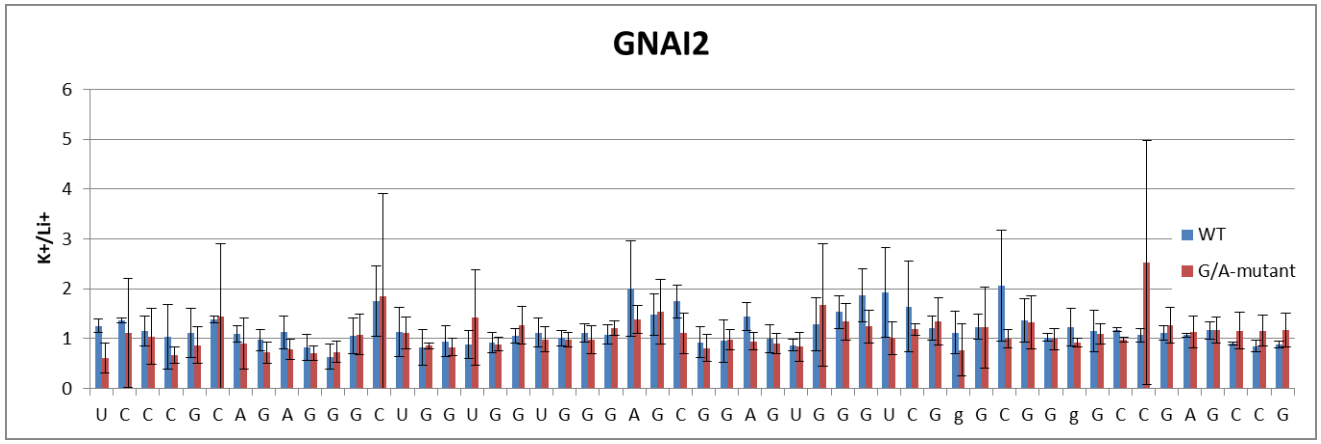
**S23**

5'-
GGGCCGCCCGGCUGCGGGGCCAGUGGCAGGAGCGCCGCGCACCGCCAGCCGCAGgGgGCGUGgGAUGgGgGCGGCCGGgGAGGgGgGCGCCCACACUGACUAGAGCC
AACCGCGCACUUCAAAAGGGUGU-3'



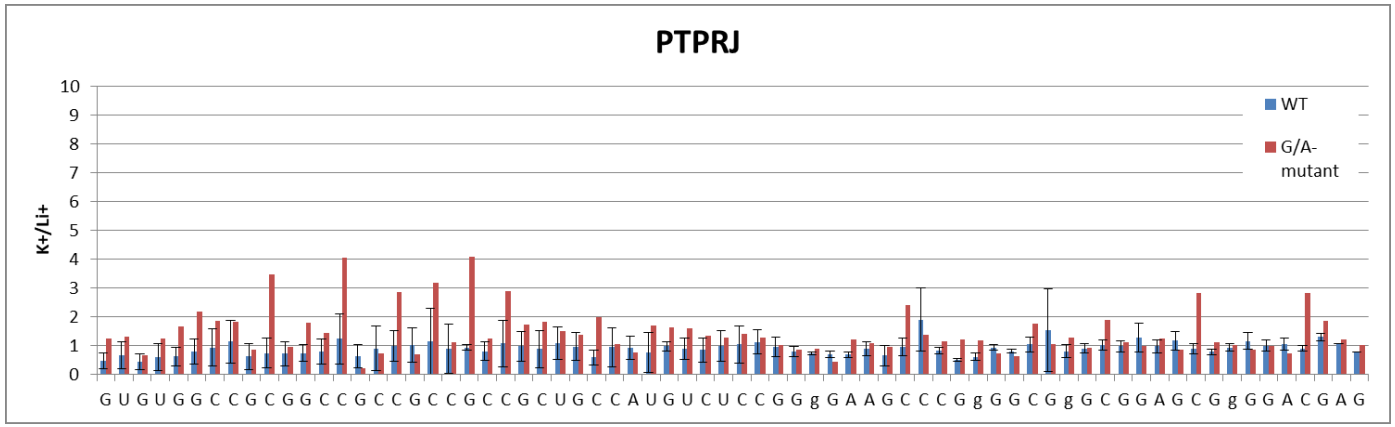

**S24**

5'-
GGGCCGACCCGAGUGCUUCCCGCAGAGGGCUGGUGGUGGGAGCGGAGUGGGUCGgGCGGgGCCGAGCCGgGCCGUGgGCCGUGUGGGGGCCGGGCGGCGGCCGGGCC
GGCGGACGGCGGG-3'



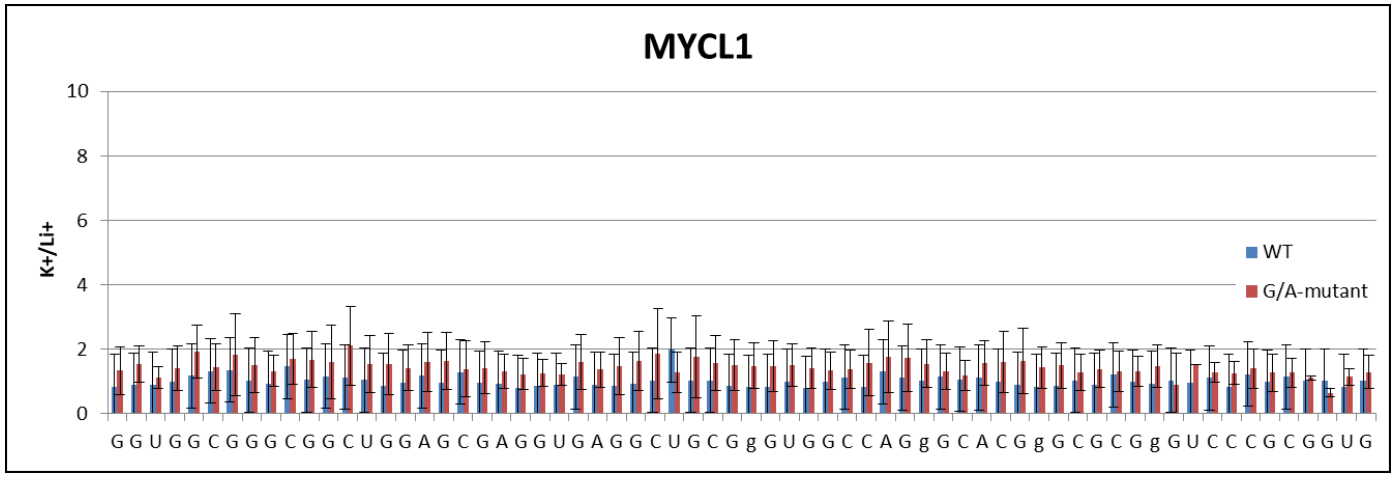
**S25**

5'-
GGGCUAGGCUCCGGCGUGUGGCCGCGGCCGCCGCCGCCGCUGCCAUGUCUCC**GGgGAAGCCCGgGGCGgGCGGAGCGgGG**ACGAGGCGGACCGGCUGGCGGAGGAGG
AGGCGAAGGAGACGGCAGGAGG-3'

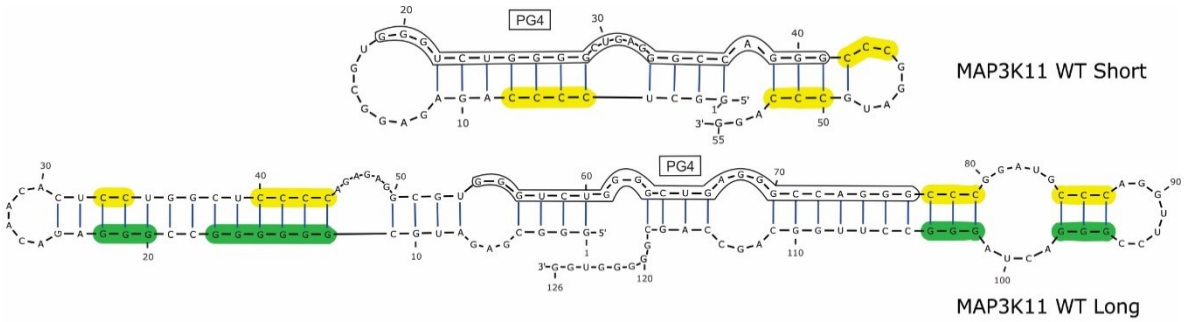



**S26**

5'-
GGGAGCCGGUCCGCUCCAGGUGGCGGGCGGCUGGAGCGAGGUGAGGCUGC**GgGUGGCCAGgGCACGgGCGCGgG**UCCCGCGGUGCGGGCUGGCUGCAGGCUGCCUUC
UGGGCACGGCGCGCCCC-3'



**S27**




**Supplementary Figure Legends**

**Supplementary figures S1 to S26.** In-line probing results of all of the PG4 candidates are shown as bar graphs of the $K^+/Li^+$ ratios. The full length sequence of each candidate is also shown. The PG4 regions are written in blue. Lowercase red guanosines are those mutated to adenines in the G/A-mutants. The lowercase green cytosines are those mutated to adenines in the C/A-mutants. For the TTYH1 PG4 candidate, the lowercase yellow cytosines are those mutated to adenines in the second C/A-mutant version. The guanines boxed in red are those involved in G4 formation.

**Supplementary figure S27**. RNAfold predicted secondary structures for both the MAP3K11 short and long transcripts. The PG4 region is boxed. Inhibitory tracks of cytosines are shown in yellow, and enhancing tracks of guanines are shown in green. In the short transcript the C-tracks are predicted to base-pair with the G-tracks of the PG4, thereby inhibiting its formation. In the longer transcript, the supplementary G-tracks base-pair with the inhibitory C-tracks allowing the G4 to form.

**ANNEXE 3 Supplementary data Article 3**

# Supplementary data

## Article 3 – The folding of 5'UTR human G-quadruplexes possessing a long central loop

**Table S1** Complete 5'-UTR RNA sequences used for in cellulo assays

**Table S2** List of DNA oligonucleotides used for synthesis of the complete 5'-UTR constructs

**Table S1** Complete 5'-UTR RNA sequences used for in cellulo assays

| **Candidate** | **Construct** | |
|---|---|---|
| HIRA | WT | 5'-<br>GAUGCGGCUGUGGUGGCGGCGGCGGCGGCCGAGCGCGGGUGGCGGCUGUGGCGGCGGAGGGGGGCGC<br>GGGCCGGCGAUGGCGCGGCGGCCCUGAGGGCGCGGGGCGGGCGGCGGCCGGAGGGCGGGUGGCGCGG<br>GAGGAAGCGGCGGCGGUGGCUCCAUGGCCCGGGCGCGCUGAGGGACCCGGCGCUCGCCUCAGCCCGG<br>CGGCGGCGGCGGCCGAACA-3' |
| | G/A-mut Central loop | 5'-<br>GAUGCGGCUGUGGUGGCGGCGGCGGCGGCCGAGCGCGGGUGGCGGCUGUGGCGGCGGAGGGGGGCGC<br>GGGCCGGCGAUGGCGCGGCGGCCCUGAGGGCGCGGGGCGGGCaaCaaCCaaAGGGCGGGUGGCGCGGGA<br>GGAAGCGGCGGCGGUGGCUCCAUGGCCCGGGCGCGCUGAGGGACCCGGCGCUCGCCUCAGCCCGGCG<br>GCGGCGGCGGCCGAACA-3' |
| | G/A-mut G-tracts | 5'-<br>GAUGCGGCUGUGGUGGCGGCGGCGGCGGCCGAGCGCGGGUGGCGGCUGUGGCGGCGGAGGGGGGCGC<br>GGGCCGGCGAUGGCGCGGCGGCCCUGAGGGCGCGaaGCGaGCGGCGGCCGGAGaGCGaGUGGCGCGGGA<br>GGAAGCGGCGGCGGUGGCUCCAUGGCCCGGGCGCGCUGAGGGACCCGGCGCUCGCCUCAGCCCGGCG<br>GCGGCGGCGGCCGAACA-3' |
| APC | WT | 5'-<br>AGAGCUAGCAGUCUUCCCACCUCCCACAAGAUGGCGGAGGGCAAGUAGCAAGGGGGCGGGGUGUGGC<br>CGCCGGAAGCCUAGCCGCUGCUCGGGGGGGACCUGCGGGCUCAGGCCCGGGAGCUGCGGACCGAGGU<br>UGGCUCGAUGCUGUUCCCAGGUACUGUUGUUGGCUGUUGGUGAGGAAGGUGAAGCACUCAGUUGCC<br>UUCUCGGGCCUCGGCGCCCCCUGCUAGCACUGACU-3' |
| | G/A-mut G-tracts | 5'-<br>AGAGCUAGCAGUCUUCCCACCUCCCACAAGAUGGCGGAGGGCAAGUAGCAAGaGaGCGaaGUGUGGCC<br>GCCGGAAGCCUAGCCGCUGCUCGaGaGaGACCUGCGGGCUCAGGCCCGGGAGCUGCGGACCGAGGUUG<br>GCUCGAUGCUGUUCCCAGGUACUGUUGUUGGCUGUUGGUGAGGAAGGUGAAGCACUCAGUUGCCUU<br>CUCGGGCCUCGGCGCCCCCUGCUAGCACUGACU-3' |
| TOM1L2 | WT | 5'-<br>AGAGCUAGCAGAGACGCGGCAAGGGGGCGGGGCCAAAGGCCCUAAGCUCGGCGUUCCAGAGAGUGGG<br>GAGGGGGCAAGUGUCAGUCAGGACGGGAGUCCGGCGGGUUACAGCGGAGGCCUAGGUGGCAGACAG<br>GGGGCCCGGGCCGCUGCGUGUUGUCCACCCAAGGCUAGCACUGACU-3' |
| | G/A-mut G-tracts | 5'-<br>AGAGCUAGCAGAGACGCGGCAAGaGaGCGaaGCCAAAGGCCCUAAGCUCGGCGUUCCAGAGAGUGaaGA<br>GaGaGCAAGUGUCAGUCAGGACGGGAGUCCGGCGGGUUACAGCGGAGGCCUAGGUGGCAGACAGGGG<br>GCCCGGGCCGCUGCGUGUUGUCCACCCAAGGCUAGCACUGACU-3' |

| | | |
|---|---|---|
| MDS1 | WT | 5’-<br>GCUAGCGAUUGCCAUCUGACAAGAUCUCCAAAUCAAAGUGAUAAAUCGCUCCAAACUUUUUUUGGCG<br>GCGCUGAGAUGUUGGAGGGGCGUCUAGCGCGCAUGUGCGAAGGUGUCCAAACUGACAAUGCUGGAG<br>AGAUAGCGAGUGUGGAUUGAGAGAAAGGGAGAGAGGGAGGGAGAGAGAGUGAAAGAAGAAAAUACA<br>GAGAGUGAGUGUGUGGAAGAGAGAGAGAAACAGGAGAGAAACAGGAGGGAGGGAGAGAGAGAGAGA<br>GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGACAGGAGAGAGAGGGAGGGAGCGAG<br>AGGGAGAGCAAAAGAAGGAAAGGAUCCAAGAAAAAAAAGCCCCAACCACACACCAGCGGCUGCAGGA<br>CUGGGCACAGCGCUAGC-3' |
| | G/A-mut G-tracts | 5’-<br>GCUAGCGAUUGCCAUCUGACAAGAUCUCCAAAUCAAAGUGAUAAAUCGCUCCAAACUUUUUUUGGCG<br>GCGCUGAGAUGUUGGAGGGGCGUCUAGCGCGCAUGUGCGAAGGUGUCCAAACUGACAAUGCUGGAG<br>AGAUAGCGAGUGUGGAUUGAGAGAAAGaGAGAGAGaGAGaGAGAGAGAGUGAAAGAAGAAAAUACAG<br>AGAGUGAGUGUGUGGAAGAGAGAGAGAAACAGGAGAGAAACAGGAGaGAGaGAGAGAGAGAGAGAG<br>AGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGACAGGAGAGAGAGAGAGAGAGCGAGA<br>GGGAGAGCAAAAGAAGGAAAGGAUCCAAGAAAAAAAAGCCCCAACCACACACCAGCGGCUGCAGGAC<br>UGGGCACAGCGCUAGC-3' |
| LRRC37A3 | WT | 5’-<br>GCUAGCUGGGAUUAUAGGCGUGAGCCACUGCACCUGGCACAGGCUGAAGUGCAAUGUUGUGAUCUCG<br>GCUCACUGCAACCUCUGCCUCCCAGGUUCAAGCGAUUCUCCUGCUUUGGCCUCCUGAGUAGCUGGGA<br>UUGCAGGUCAGUUGCUCUCCCUGGAAGGAAGAGUGUUCUCGGAUUUCACCUUAAAGGAGGAAGGCU<br>GCCAGAACUGAACUAGCACUUCUGAAUAUCCUGAGGCGAGGUCCGGUGACUUCCUUGGGAAGCUCUG<br>CCGCGCCCCCAUCCCACCCUACCCCACCCUACCCCACCACAGCAGGCGCUGGAGUCCUGGGACCACCA<br>GGAUCUGAGGCCCAAAUCCUUCCUCACUAAGGGGAGGAGAGGGGUGCUCCGGCAGGGCAGGAUGGGA<br>AGGCGUGCUUGGGCGGGAUUGUGACAUAAGAGUGCCCUGGUGACAUGGAGCAGAUCUGUGGCAUAA<br>AUAAAGGUGUCAUAAAGACAGGGCGGGACUCAUGCUUACAAGGGGCACGAGCGUCUCGGAGCUGCCA<br>GAGCUAGC-3' |
| | G/A-mut G-tracts | 5’-<br>GCUAGCUGGGAUUAUAGGCGUGAGCCACUGCACCUGGCACAGGCUGAAGUGCAAUGUUGUGAUCUCG<br>GCUCACUGCAACCUCUGCCUCCCAGGUUCAAGCGAUUCUCCUGCUUUGGCCUCCUGAGUAGCUGGGA<br>UUGCAGGUCAGUUGCUCUCCCUGGAAGGAAGAGUGUUCUCGGAUUUCACCUUAAAGGAGGAAGGCU<br>GCCAGAACUGAACUAGCACUUCUGAAUAUCCUGAGGCGAGGUCCGGUGACUUCCUUGGGAAGCUCUG<br>CCGCGCCCCCAUCCCACCCUACCCCACCCUACCCCACCACAGCAGGCGCUGGAGUCCUGGGACCACCA<br>GGAUCUGAGGCCCAAAUCCUUCCUCACUAAGGGGAGGAGAGGGGUGCUCCGGCAGGGCAGGAUGGGA<br>AGGCGUGCUUGaGCGaGAUUGUGACAUAAGAGUGCCCUGGUGACAUGGAGCAGAUCUGUGGCAUAAA<br>UAAAGGUGUCAUAAAGACAGaGCGaGACUCAUGCUUACAAGGGGCACGAGCGUCUCGGAGCUGCCAG<br>AGCUAGC-3' |

Mutated nucleotides are in lowercase

**Table S2** List of DNA oligonucleotides used for synthesis of the complete 5'-UTR constructs

| **Candidate** | **Name of oligo** | |
|---|---|---|
| | Seq 5UTR pRLTK | 5’-GGTCTTACTGACATCCAC-3' |
| HIRA | Hira 1 | 5’-GATGCGGCTGTGGTGGCGGCGGCGGCGGCCGAGCGCGGGTGGCGGCTGTGGCGGCGGAGGGGGGGCGCGGGCCGGCGATGGCGCGG-3' |
| | Hira 2 wt | 5’-GCCGCTTCCTCCCGCGCCACCCGCCCTCCGGCCGCCGCCCGCCCCGCGCCCTCAGGGCCGCCGCGCCATCGCCGGC-3' |
| | Hira 2 mut loop | 5’-GCCGCTTCCUCCCGCGCCACCCGCCCTTTGGTTGTTGCCCGCCCCGCGCCCTCAGGGCCGCCGCGCCATCGCCGGC-3' |
| | Hira 2 mut tracts | 5’-GCCGCTTCCTCCCGCGCCACTCGCTCTCCGGCCGCCGCTCGCTTCGCGCCCTCAGGGCCGCCGCGCCATCGCCGGC-3' |
| | Hira 3 | 5’-TGTTCGGCCGCCGCCGCCGCCGGGCTGAGGCGAGCGCCGGGTCCCTCAGCGCGCCCGGGCCATGGAGCCACCGCCGCCGCTTCCTCCCGCGCC-3' |
| | Hira Fwd NheI | 5’-ATTGCTTAAAGCTAGCGATGCGGCTGTGGTGGCGGC-3' |
| | Hira Rev NheI | 5’-ATTGCTTAAAGCTAGCTGTTCGGCCGCCGCCGCCGC-3' |
| APC | APC fwd Wt_1 | 5’-CCTCCCACAAGATGGCGGAGGGCAAGTAGCAAGGGGGCGGGGTGTGGCCGCCGGAAGCCTAGCCGCTGCTCGGGGGGGACCTGCGGGCTCAGGCCCGGG-3’ |
| | APC fwd G/A-mut_1 | 5’-CCTCCCACAAGATGGCGGAGGGCAAGTAGCAAGAGAGCGAAGTGTGGCCGCCGGAAGCCTAGCCGCTGCTCGAGAGAGACCTGCGGGCTCAGGCCCGGG-3’ |
| | APC rev_1 | 5'-CTGAGTGCTTCACCTTCCTCACCAACAGCCAACAACAGTACCTGGGAACAGCATCGAGCCAACCTCGGTCCGCAGCTCCCGGGCCTGAGCCCGCAGGTC-3' |
| | Nhe1_APC fwd_2 | 5’-AGAGCTAGCAGTCTTCCCACCTCCCACAAGATGGCGGAGGG-3’ |

| | | |
|---|---|---|
| | Nhe1_APC rev_2 | 5'-AGTCAGTGCTAGCAGGGGGCGCCGAGGCCCGAGAAGGCAACTGAGTGCTTCACCTTCC-3' |
| TOM1L2 | TOM1L2 fwd wt | 5'-AGAGCTAGCAGAGACGCGGCAAGGGGGCGGGGCCAAAGGCCCTAAGCTCGGCGTTCCAGAGAGTGGGGAGGGGGCAAGTGTCAGTCAGGACGGGAGTCCG-3' |
| | TOM1L2 fwd G/A-mut | 5'-AGAGCTAGCAGAGACGCGGCAAGAGAGCGAAGCCAAAGGCCCTAAGCTCGGCGTTCCAGAGAGTGAAGAGAGAGCAAGTGTCAGTCAGGACGGGAGTCCG-3' |
| | TOM1L2 rev | 5'-AGTCAGTGCTAGCCTTGGGTGGACAACACGCAGCGGCCCGGGCCCCCTGTCTGCCACCTAGGCCTCCGCTGTAACCCGCCGGACTCCCGTCCTGACTGAC-3' |

**ANNEXE 4 Supplementary data Article 4**

# Supplementary data

## Article 4 – G-quadruplexes formation in the 5'UTRs of mRNAs associated with colorectal cancer pathways

**Supplementary tables S1–S5**

**S1 Table.** Sequences, positions in the 5'UTR and lengths of all candidates and their respective full-length 5'UTRs.

**S2 Table**. Comparison of the prediction methods.

**S3 Table.** UTRref, RefSeq and Gene-ontology Identification numbers of all candidates.

**S4 Table.** Oligonucleotide sequences used for PCR-filling prior to in vitro transcription.

**S5 Table**. Oligonucleotide sequences used for PCR filling prior to cloning.

**Supplementary figures S1–S4**

**S1 Fig**. *In-line* probing gels and $K^+/Li^+$ ratio quantification of the candidates.

**S2 Fig.** NMM assay of all candidates.

**S3 Fig**. *In cellulo* luciferase assay in HEK293 cells.

**S4 Fig.** *In cellulo* luciferase assay in colorectal cancer cell lines.

## Supplementary tables

### Table S1 Sequences, positions in the 5'UTR and lengths of all candidates and their respective full-length 5'UTRs

| | | Sequences 5'-3' orientation | Position in the 5'UTR | Length (nts) |
|---|---|---|---|---|
| **ACVR1C** | WT | **GGG**CCGCCCGGCUGCGGGGCCAGUGGCAGGAGCGCCGCGCACCGCCAGCCGCAGGGGGCGUGGGAUGGGGCGGCCGGGGAGGGGGG CGCCCACACUGACUAGAGCCAACCGCGCACUUCAAAAGGGUGU | 6-132 (+3) | 130 |
| | G/A-mutant | **GGG**CCGCCCGGCUGCGGGGCCAGUGGCAGGAGCGCCGCGCACCGCCAGCCGCAGAGAGCGUGAGAUGAGAGCGGCCGGAGAGGAGAG CGCCCACAGUGACUAGAGCCAACCGCGCACUUCAAAAGGGUGU | | 130 |
| | UTR | GCUCCCCGCCCGGCUGCGGGGCCAGUGGCAGGAGCGCCGCGCACCGCCAGCCGCAGGGGGCGUGGGAUGGGGGCGGCCGGGGAGGGG GGCGCCCACACUGACUAGAGCCAACCGCGCACUUCAAAAGGGUGUCGGUGCCGCGCUCCCCUCCCGCGGCCCGGGAACUUCAAAGCG GGCCGUGCUGCCCCGGCUGCCUCGCUCUGCUCUGGGGCCUCGCAGCCCCGGCGCGGCCGCCUGGUGGCG | | 243 |
| **AIFM2** | WT | GGGAAGACGACCAAGCGGGAGCGGGAGCGGGAGCGGGAGCCGGAGCGAGAGCGCGCGGGCG | 131-191 | 61 |
| | G/A-mutant | GGGAAGACGACCAAGCGAGAGCGAGAGCGAGAGCGAGAGCCGAAGCGAGAGCGCGCGGGCG | | 61 |
| | UTR | GGCCGUUCGGAGACCAGCCCCAGCGUGCCAGGACCGUUUCCGGGGCCUGGCCGGGGCGUUGCCGCGGGGUCGGGGACCAGCACGAGU GCUGAGUCACGCCCCGCCCGGGAGCGCCUCGGGUCAGUAACUCGGGAAGACGACCAAGCGGGAGCGGGAGCGGGAGCGGGAGCCGGA GCGAGAGCGCGCGGGCGCGGCCGACAGUGCCUGAUUUGAG | | 214 |
| **APC** | WT | GGGCAAGUAGCAAGGGGGCGGGGUGUGGCCGCCGGAAGCCUAGCCGCUGCUCGGGGGGGACCUGCGGGCUCAGG | 14-87 | 74 |
| | G/A-mutant | GGGCAAGUAGCAAGAGAGCGAAGUGUGGCCGCCGGAAGCCUAGCCGCUGCUCGAGAGAGACCUGCGGGCUCAGG | | 74 |
| | UTR | ACAAGAUGGCGGAGGGCAAGUAGCAAGGGGGCGGGGUGUGGCCGCCGGAAGCCUAGCCGCUGCUCGGGGGGGACCUGCGGGCUCAGG CCCGGGAGCUGCGGACCGAGGUUGGCUCGAUGCUGUUCCCAGGUACUGUUGUUGGCUGUUGGUGAGGAAGGUGAAGCACUCAGUUGC CUUCUCGGGCCUCGGCGCCCCCUAUGUACGCCUCCCUGGGCUCGGGUCCGGUCGCCCCUUUGCCCGCUUCUGUACCACCCUCAGUUC UCGGGUCCUGGAGCACCGGCGGCAGCAGGAGCUGCGUCCGGCAGGAGACGAAGAGCCCGGGCGGCGCUCGUACUUCUGGCCACUGGG CGAGCGUCUGGCAGGUCCAAGGGUAGCCAAGG | | 380 |
| **APPL1** | WT | **GG**GCUCGGCGCCUGGAGAAGGCUGUGCGGGCGGGGACGGCUGCAGCCCUUGCCGGAGAGGGCGGGCCGGGGUCAGCUGCGGCGGGCG GGCCGGCGCGGGGAGCUGUGGGCGGCAGCUGCGUCUCCUGCCACCGCCCUCCCUCCGCCACG | 1-147 (+2) | 149 |
| | G/A-mutant | **GG**GCUCGGCGCCUGGAGAAGGCUGUGCGGGCGGGGACGGCUGCAGCCCUUGCCGGAGAGAGCGAGCCGGAGUCAGCUGCGGCGAGCG AGCCGGCGCGAGGAGCUGUGAGCGGCAGCUGCGUCUCCUGCCACCGCCCUCCCUCCGCCACG | | 149 |
| | UTR | GCUCGGCGCCUGGAGAAGGCUGUGCGGGCGGGGACGGCUGCAGCCCUUGCCGGAGAGGGCGGGCCGGGGUCAGCUGCGGCGGGCGGG CCGGCGCGGGGAGCUGUGGGCGGCAGCUGCGUCUCCUGCCACCGCCCUCCCUCCGCCACG | | 147 |

| | | | | |
|---|---|---|---|---|
| **BAD** | WT | GGGUCAGGGGCCUCGAGAUCGGGCUUGGGGUGAGACCUGUGCGCCGUCACCACGGGCGGGGCGCGGGCCUGGGUCCACCGGGGUUCUG AGGGGAGACUGAGGUCCUGAGCCGACAGCCUCAGCUCCCUGCCA | 49-176 | 131 |
| | G/A-mutant | GGGUCAGGAGCCUCGAGAUCGAGCUUGAGGUGAGACCUGUGCGCCGUCACCACGAGCGGAGCGAGGCCUGAGUCCACCGGAGUUCUG AGGGGAGACUGAGGUCCUGAGCCGACAGCCUCAGCUCCCUGCCA | | 131 |
| | UTR | AACUAGGGCCCGGAGCCCGGGGUGCUGGAGGGAGGCGGCAGGCCCGGGUCAGGGGCCUCGAGAUCGGGCUUGGGGUGAGACCUGUGC GCCGUCACCACGGGCGGGGCGGGGCCUGGGUCCACCGGGGUUCUGAGGGGAGACUGAGGUCCUGAGCCGACAGCCUCAGCUCCCUGC CAGGCCAGACCCGGCAGACAGAUGAGGGCCCAGGAGGCCUGGCGGGCCUGGGGGCGCUACGGUGGGAGAGGAAGCCAGGGGUACCUG CCUCUGCCUUCCAGGGCCACCGUUGGCCCCAGCUGUGCCUUGACUACGUAACAUCUUGUCCUCACAGCCCAGAGC | | 336 |
| **BAG-1** | WT | **GG**GCAGGCCGGGGCGGGGCUGGGAAGUAGUCGGGCGGGGUUGUGAGACGCCGCGCUCAGCUUCCAUCGCUGGGCGGUCAACAA | 1-81 (+2) | 83 |
| | G/A-mutant | **GG**GCAGGCCGAGACGAGACUGAGAAGUAGUCGAGCGAGGUUGUGAGACGCCGCGCUCAGCUUCCAUCGCUGGGCGGUCAACAA | | 83 |
| | UTR | GCAGGCCGGGGCGGGGCUGGGAAGUAGUCGGGCGGGGUUGUGAGACGCCGCGCUCAGCUUCCAUCGCUGGGCGGUCAACAAGUGCGG GC | | 87 |
| **BAG-5** | WT | GGCGCGGACGCCGGAGGAGGUGUCCCCGGGUUUAGGGGUGUUCGGCCAGGGGCGGGGCUGCCGGGCCCGGGCGACUGCCGGAGCUGC GGAAGUCGUGGAAGCGUCGGCGACGCAUCGCGCG | 61-181 | 121 |
| | G/A-mutant | GGCGCGGACGCCGGAGGAGGUGUCCCCGAGUUUAGGAGUGUUCGGCCAAGAGCAGAGCUGCCGAGCCCGAGCGACUGCCGGAGCUGC GGAAGUCGUGGAAGCGUCGGCGACGCAUCGCGCG | | 121 |
| | UTR | CGGUCACUCCAGGCUGCCUCGCUGCGCCGCUGGCGGUGGGCGUGGCGCUGCCGGCGGCUAGGCGCGGACGCCGGAGGAGGUGUCCCC GGGUUUAGGGGUGUUCGGCCAGGGGCGGGGCUGCCGGGCCCGGGCGACUGCCGGAGCUGCGGAAGUCGUGGAAGCGUCGGCGACGCA UCGCGCGAUGGCGCGGGCGGGACAGUGCUUGUGAAACUGAACACAACAAAAGU | | 227 |
| **BCL-2** | WT | **GG**GCAAAGCACAUCCAAUAAAAUAGCUGGAUUAUAACUCCUCUUCUUUCUCUGGGGGCCGUGGGGUGGGAGCUGGGGCGAGAGGUGC CGUUGGCCCCCGUUGCUUUUCCUCUGGGAAGG | 377-493 (+2) | 119 |
| | G/A-mutant | **GG**GCAAAGCACAUCCAAUAAAAUAGCUGGAUUAUAACUCCUCUUCUUUCUCUGAGAGCCGUGAGAUGAGAGCUGGAGCGAGAGGUGC CGUUGGCCCCCGUUGCUUUUCCUCUGGGAAGG | | 119 |
| | UTR | UUUCUGUGAAGCAGAAGUCUGGGAAUCGAUCUGGAAAUCCUCCUAAUUUUUACUCCCCUCUCCCCCGCGACUCCUGAUUCAUUGGGAAG UUUCAAAUCAGCUAUAACUGGAGAGUGCUGAAGAUUGAUGGGAUCGUUGCCUUAUGCAUUUGUUUUGGUUUUACAAAAAGGAAACUU GACAGAGGAUCAUGCUGUACUUAAAAAAUACAACAUCACAGAGGAAGUAGACUGAUAUUAACAAUACUUACUAAUAAUAACGUGCCU CAUGAAAUAAAGAUCCGAAAGGAAUUGGAAUAAAAAUUUCCUGCAUCUCAUGCCAAGGGGGAAACACCAGAAUCAAGUGUUCCGCGU GAUUGAAGACACCCCCUCGUCCAAGAAUGCAAAGCACAUCCAAUAAAAUAGCUGGAUUAUAACUCCUCUUCUUUCUCUGGGGGCCGU GGGGUGGGAGCUGGGGCGAGAGGUGCCGUUGGCCCCCGUUGCUUUUCCUCUGGGAAGG | | 493 |
| **BCL-9L** | WT | **GGG**CGCUCGCUCGCUCUGCCUCUCCGCCCGGGCUCUGCCGAAGGGGGCGGGGUGGGGGGUGCAGGGCGGGGGGAGGGGAGGGCUCCUGC AUUCUUGCGGUCGGGGAGGAAUCCGAGCCAGCGUUACUGG | 234-357 (+3) | 127 |
| | G/A-mutant | **GGG**CGCUCGCUCGCUCUGCCUCUCCGCCCGAGCUCUGCCGAAGAGAGCGAGAUGAGAGUGCAGAGCAGAGAGAGAGAAGGCUCCUGC AUUCUUGCGGUCGAGGAGGAAUCCGAGCCAGCGUUACUGG | | 127 |

| | | | | |
|---|---|---|---|---|
| | UTR | CGUGCGUGUCUUGUCUCCCUGUCCACGUGUGAGCUGUGAGUGUGUGUGAGUCAGAGUUCGGGUGUCUGUGGGUCUCUGAGCCUCUGC<br>UGGCAGCACCCGGGGCUCGCCAAGCUCUUGCCGGCUGGCGCGCGCCCAGCCCCUGGCGGGACUUGUCCGUGUGUCUGUCCGCGCGCG<br>GGGCCUGGAGAAGCACGCUGCAGUCCUGUCCGCCUCCCGCUCGCUGCCUCGCUGGCUGUCGCUCGCUCGCUCUGCCUCUCCGCCCGG<br>GCUCUGCCGAAGGGGGCGGGGUGGGGGUGCAGGGCGGGGGGAGGGGAGGCUCCUGCAUUCUUGCGGUCGGGGAGGAAUCCGAGCCAG<br>CGUUACUGGUCUCCAGAAGAGCCCAGCUGCAGCCCCGGGGCCCCGCCAGGCCUCUCGCUGCCCGCCCGGGCCUGCUGGGAUGGGCAC<br>GGGCUAGGCCUCGAGCUGGGGACGGGGCGGGGCGUGGCCCCAACCCCGGGCCCCCUCCACGGCUGGAGCGCUCUGGGGGUGGGGCAC<br>GAGGGGUACCCCACCCUGGGUGAGGGGCGCGUCUGGGAGCAGGAAUCCCUCAGGGGGCCAGGGGAGACCUCACAGCCGCCCCACACG<br>GCACCUUUGCUACCUAGCCUUUUAGUGAAUUCUGUCUCUGCCGCCCGUCGGGGCGGAGGCUUGCUGGAGACUGCAAGCCCCCUGAGG<br>GCAAGGUGCGGGAGGGAUGGGGACAGGGCUGGCCUCCAGGAUCGAGACCCCAUCUCGUAAUCCUCUUUUUCCAGUGCCACCCCACCC<br>UAUCCAGCUCCCUAGUCCCAGGGCUGUUGGGUCCUCCCUUCUCCCCCUCCCCCUCUGACACCCCCUCCCCAAGUCACGAGUUUUCUCU<br>UUGGGGCUUGUUGCUGCAGUCCGUGCUCCAGUACCGAGUACCUGGCUGGGCCCUGGGCACGCACAGGGGCCGUAGCCCCACUGUGUG<br>UGGGAGCC | | 965 |
| **BMPR1A** | WT | GGCAGGAGCGAGGAGGGAGGAGGCCAAGGGCGGGCAGGAAGGCUUAGGC | 58-107 | 50 |
| | G/A-mutant | GGCAGGAGCGAAGAGAGAAGAGAGCCAAGAGCGAGCAGGAAAGCUUAGGC | | 50 |
| | UTR | GCGGCCGCUGCAGAGAUUGGAAUCCGCCUGCCGGGCUUGGCGAAGGAGAAGGGAGGAGGCAGGAGCGAGGAGGGAGGAGGGCCAAGG<br>GCGGGCAGGAAGGCUUAGGCUCGGCGCGUCCGUCCGCGCGCGGCGAAGAUCGCACGGCCCGAUCGAGGGGCGACCGGGUCGGGGCCG<br>CUGCACGCCAAGGGCGAAGGCCGAUUCGGGCCCCACUUCGCCCCGGCGGCUCGCCGCGCCCACCCGCUCCGCGCCGAGGGCUGGAGG<br>AUGCGUUCCCUGGGGUCCGGACUUAUGAAAAUAUGCAUCAGUUUAAUACUGUCUUGGAAUUCAUGAGAUGGAAGCAUAGGUCAAAGC<br>UGUUUGGAGAAAAUCAGAAGUACAGUUUUAUCUAGCCACAUCUUGGAGGAGUCGUAAGAAAGCAGUGGGAGUUGAAGUCAUUGUCAA<br>GUGCUUGCGAUCUUUUACAAGAAAAUCUCACUGAAUGAUAGUCAUUUAAAUUGGUGAAGUAGCAAGACCAAUUAUUAAAGGUGACAG<br>UACACAGGAAACAUUACAAUUGAACA | | 548 |
| **BOK** | WT | **GGG**AAGAGCGCGGGAAGCCCCGUGGACCUGGCGCUCCCGGCUCGGGCGUGGACGGGGCGGGCGCCCGGGCGGGCGCGCGUCCUCGC<br>GGGUCUGAAUGGAAGGGUCGAGGUCGUCGUCGGCGGC | 41-161 (+3) | 124 |
| | G/A-mutant | **GGG**AAGAGCGCGGGAAGCCCCGUGGACCUGGCGCUCCCGGCUCGGGCGUGGACGAGACGAGCGCCGAGACGAGGCGCGCGUCCUCGC<br>GGGUCUGAAUGGAAGGGUCGAGGUCGUCGUCGGCGGC | | 124 |
| | UTR | CUCGCUGCCCAGGCCCCCGACGCCGCGGCAGGAGCCCCCCAAGAGCGCGGGAAGCCCCGUGGACCUGGCGCUCCCGGCUCGGGCGUG<br>GACGGGGCGGGCGCCGGGGCGGGGCGCGCGUCCUCGCGGGUCUGAAUGGAAGGGUCGAGGUCGUCGUCGGCGGCGAGCAGAUCCUGA<br>AGCCAGAACUCCACCCCGGCGCCCGCGCCAUGCGGCGGGAGAGGUGCGGCGCCCCCCACCCGCGUCGCCGCC | | 246 |
| **CASP6** | WT | **GG**CCGAGGGCGGGGCCGGGCCCGGGAGCCUGUGGCUUCAGGAAGAGGAGGGCAAGGUGUCUGGCUGCGCGUUUGG | 1-73 (+2) | 75 |
| | G/A-mutant | **GG**CCGAGAGCGAAGCCGGGCCCGGGAGCCUGUGGCUUCAGGAAGAGGAGGGCAAGGUGUCUGGCUGCGCGUUUGG | | 75 |
| | UTR | CCGAGGGCGGGGCCGGGCCCGGGAGCCUGUGGCUUCAGGAAGAGGAGGGCAAGGUGUCUGGCUGCGCGUUUGGCUGCA | | 78 |
| **CASP8AP2** | WT | **GGG**AAAGGAACCGGGUUGUCUUGGGCCGGCAGGGCGGGUAAGUUGUCGUAGGGGCCCGGUCCGUGAGGGACUGCUAAGGAAG | 1-80 (+3) | 83 |
| | G/A-mutant | **GGG**AAAGGAACCGAGUUGUCUUGAGCCGAGCAGAGCGAGUAAGUUGUCGUAGAGGCCCGGUCCGUGAGGGACUGCUAAGGAAG | | 83 |
| | UTR | AAAGGAACCGGGUUGUCUUGGGCCGGGCAGGGCGGGUAAGUUGUCGUAGGGGCCCGGUCCGUGAGGGACUGCUAAGGAAGAGGCUGC<br>AUGGCGCGGUAGUCCCCCGAGUGGAGGUCGGCUGCCCCUGGGAAACCAGAGAGUCGGAGGGAGUCCAUCUGGAGCGGCCAAGUAGGU<br>CGGGGAAGGGCCGCGCCUGACGUCUGGCCGCAGGUGGUGACUCUCAAAAGGAAAUAGGAUC | | 235 |
| **CASP9** | WT | GGCCCUGGGGCGGGGCGGGUCCUGGGGACUGGGCGGGCGGCCGAGGCCCGGAAGCGGACUGAGGCGGCCUGGAGUC | 1-78 | 78 |

| | | | | |
|---|---|---|---|---|
| | G/A-mutant | GGCCCUGGAGCGAGAGCGAGUCCUGAGGACUGAGGCGAGCGGCCGAGGCCCGGAAGCGGACUGAGGCGGCCUGGAGUC | | 78 |
| | UTR | GGCCCUGGGGCGGGGGCGGGUCCUGGGGACUGGGGCGGGCGGCCGAGGCCCGGAAGCGGACUGAGGCGGCCUGGAGUCUUAGUUGGC UACUCGCC | | 95 |
| **FZD10** | WT | GGGGGCGCUGUGCGCAGCGCUCGGGCCAGGCCGGGCGGCAUGGGCGGGGCCCGAGCAGGGGUGG | 305-370 | 66 |
| | G/A-mutant | GGGGGCGCUGUGCGCAGCGCUCGAGCCAGGCCGAGCGAGCAUGAGCGAGAGCCCGAGCAGGGGUGG | | 66 |
| | UTR | UCGAAACAGCUGCCGGCUGGUCCCGGCCGAGGCCGGCGCAGGGAGGGAGGAGCCGCCCGGGCUGUGGGGGCGCCGCGAGCUGGGCCG GCCUCGGUGUGCCCGCGCCGCCAGCCCGCUCCAGACGCGCCACCUGGGCGCUCCAAGAAGAGGCCGAAGUUUGCCGCGGCCGUGAGU UGGAGCUCGCGCCGGGCCGCUGCGCCGGGAGCUCCGGGGGCUUCCCUCGCUUCCCGGUAUUGUUUGCAAACUUUGCUGCUCUCCGCC GCGGCCCCCAACUCGGCGGACGCCGGGCGCGGAGAGCCGAGCCGGGGGCGCUGUGCGCAGCGCUCGGGCCAGGCCGGGCGGGCAUGG GCGGGGGCCCGAGCAGGGGUGGAGAGCCGGGGCCAGCAGCAGCCCGUGCCCGGGAGCGGCGGCGCUGAGGGGCGCGGAGCUCCCCGC GAGGACACGUCCAACGCCAGC | | 456 |
| **FZD2** | WT | GGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGUCUCCGGGUUGGGGGCGGGGCGGGGGGGCGCCAAGGAGCC GGGUGGGGGGGCGGCGGCCAGCGCUAGCCACCAUGACU | 55-114 | 124 |
| | G/A-mutant | GGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGUCUCCGGGUUGGAAGCGAGAGCGAAGGAAGCGCCAAGGAGCC GGGUGGGGGGGCGGCGGCCAGCGCUAGCCACCAUGACU | | 124 |
| | UTR | CGAGUAAAGUUUGCAAAGAGGCGCGGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGUCUCCGGGUUGGGGGCGG GGGCGGGGGGGGCGCCAAGGAGCCGGGUGGGGGGCGGCGGCCAGC | | 132 |
| **MAP2K1** | WT | GGGCAGCCUUUCGGCUCUCUGCGCGCGAAGCCGAGUCCCGGGCGGGUGGGGCGGGGGUCCACUGAGACCGCUACCGGCC | 247-313 (+2) | 79 |
| | G/A-mutant | GGGCAGCCUUUCGGCUCUCUGCGCGCGAAGCCGAGUCCCGAGCGAGUGAGACGAGAGUCCACUGAGACCGCUACCGGCC | | 79 |
| | UTR | AGGCGAGGCUUCCCCUUCCCCGCCCCUCCCCCGGCCUCCAGUCCCUCCCAGGGCCGCUUCGCAGAGCGGCUAGGAGCACGGCGGCGG CGGCACUUUCCCCGGCAGGAGCUGGAGCUGGGCUCUGGUGCGCGCGCGGCUGUGCCGCCCGAGCCGGAGGGACUGGUUGGUUGAGAG AGAGAGAGGAAGGGAAUCCCGGGCUGCCGAACCGCACGUUCAGCCCGCUCCGCUCCUGCAGGGCAGCCUUUCGGCUCUCUGCGCGCG AAGCCGAGUCCCGGGCGGGUGGGGCGGGGGUCCACUGAGACCGCUACCGGCCCCUCGGCGCUGACGGGACCGCGCGGGGCGCACCCG CUGAAGGCAGCCCCGGGGCCCGCGGCCCGGACUUGGUCCUGCGCAGCGGGCGCGGGGCAGCGCAGCGGGAGGAAGCGAGAGGUGCUG CCCUCCCCCCGGAGUUGGAAGCGCGUUACCCGGGUCCAAA | | 475 |
| **MAPK3** | WT | **GG**GCGGGUGACAGGCAGGCGGGAAGGGGCGGGGCCUCGGGCGGGGCCGCCGUGGGGAGGAGGGCGGUGGGAGGGGAGGAGUGGAG | 18-100 | 85 |
| | G/A | **GG**GCGGGUGACAGGCAGGCGGGAAGGGGCGGGGCCUCGGGCGGGGCCGCCGUGAGGAGGAGAGCGGUGAGAGGAGAGGAGUGGAG | | |
| | 1stG/A | **GG**GCGGGUGACAGGCAGGCGAGAAGAGACGGAGCCUCGAGCGAGGCCGCCGUGGGGAGGAGGGCGGUGGGAGGGGAGGAGUGGAG | | |
| | G/A mut all | **GG**GCGGGUGACAGGCAGGCGAGAAGAGACGGAGCCUCGAGCGAGGCCGCCGUGAGGAGGAGAGCGGUGAGAGGAGAGGAGUGGAG | | 85 |
| | UTR | CUGGCGCGCGCGGCCCUGCGGGUGACAGGCAGGCGGGAAGGGGCGGGGCCUCGGGCGGGGCCGCCGUGGGGAGGAGGGCGGUGGGAG GGGAGGAGUGGAG | | 100 |
| **PIK3R1** | WT | GGCGGUGGCCCGGACGCACUGCCGGGCGGGCGUGGGCGGAGGGACGAGCCGAGCCGAG | 44-103 | 60 |

| | | | | |
|---|---|---|---|---|
| | G/A-mutant | GGCGAUGACCCGAACGCACUGCCGAGCGAAGCGUGAAGCGGAGAGACGAGCCGAGCCGAG | | 60 |
| | UTR | AGCGAAAUCCAGUUGGCUUCUCAAUGAGGAGCCGGCAGUGAGCGGCGGUGGCCCGGACGCACUGCCGGGCGGGGCGUGGGGCGGAGG GACGAGCCGAGCCGAGCCAAGCGGAGCUGGGCCACUGUGCACGCCCGGAGGGUCCUGGCGGCGCCCCCGCUCCUGCGCGCACUCUCG GCGCCGGACACGAGCACUGCCUGCCGGGAACAGGCUGGGGGGAGGUGCGGGGGCUUGGCCCACUUGGUGGAAGAACAGCUUUGGGGA UUUUUUUUUUUUCAUUGUCGGAUACAGGCAUUUCAAAGGGAAACCGUUGAA | | 312 |
| **PIK3R3** | WT | **G**GCCUCAGCGGGUGGGCAGCAUGGGGCGGGAGGGUGUCCCCUCCGCGCCGUUAAAAUGAAACUCUAGUGGCUGGAGUCCGGGCA | 1-84 (+1) | 85 |
| | G/A-mutant | **G**GCCUCAGCGAGUGAGCAGCAUGAGGCGAAGAGAGUGUCCCCUCCGCGCCGUUAAAAUGAAACUCUAGUGGCUGGAGUCCGGGCA | | 85 |
| | UTR | GCCUCAGCGGGUGGGCAGCAUGGGGCGGGGAGGGUGUCCCCUCCGCGCCGUUAAAAUGAAACUCUAGUGGCUGGAGUCCGGGCAGAG CUUGAGGGCAGUUGGUGCGGUCGGGUUGGUUCUUACACCCCGGCGGGAGCGCCCAGACAAGCCGAGCUGACUGGACUUCUCCGGCCG GCCCCAUUCCCGAGGCUGCGGCAGCUUCGGUUCCGAGACCGACCGGAGAGGAGCCCGAGUCCCGGCCUCUGGGGGAUUCGCUCUCUG CAGACCAGUGGGACCCCGAAACUUGAACGCAAUCUCCAGCCCCCUUUUUUGCCUUCCUUUGUCACUUGCCCGGGUUUCUCCCAACGU GUUCUUUUUUUUCCUCUUCAUUCUCCCCUCCUUCGAAGGACACAAAAGUGGCUUCCGCGGAAAGAUUUGGAGGCGGUGGGAGCUUUUC UCCCCGGAGAGCGACUGUGUAGAAAGGAUUUUUGGGAAGCCGCUUUUUAACACCUCUGCUCUCCGUCCCCCAAGCCUCUGUGUAAUC CUCUGAGGAGAAAAGCCCAUAGCUUGAAAGUUCGGGGGCAUUUUGUUGUGUUCUGUAGGAGAGAGGGGGAGGACCCUGUUCGGGUAG UUUGGCCGGACUGGUACUGGCCGUUGGAAAACCCGAAGUACAUUUCCGUGUGGAACUUUUGCAGAUAUAUAUUUUUAGAUUUUUAAA UACCAGAUAAAAAAUAUAUGCCUUCUAUAUAUCUCCUGGCGACCUGCCCCUGACAGCGCG | | 756 |
| **SMAD2** | WT | **GG**GCGCCCGGGCCGCCGGCCGGGCCCGGCCUGGGGCGGGGCGGGAAGACGGCGGCCGGGAGUGUUUUCAGUUCCGCCUCCAAUCG CCCAUUCCC | 1-94 (+2) | 96 |
| | G/A-mutant | **GG**GCGCCCGGGCCGCCGGCCGAGCCCGAGCCTGAGAGCGGAGCGAGAAGACGGCGGCCGGGAGUGUUUUCAGUUCCGCCUCCAAUCG CCCAUUCCC | | 96 |
| | UTR | GCGCCCGGGCCGCCGGCCGGGCCCGGGCCUGGGGGCGGGGCGGGAAGACGGCGGCCGGGAGUGUUUUCAGUUCCGCCUCCAAUCGCC CAUUCCCCUCUUCCCCUCCCAGCCCCCUCCAUCCCAUCGGAAGAGGAAGGAACAAAAGGUCCCGGACCCCCCGGAUCUGACGGGGCG GGACCUGGCGCCACCUUGCAGGUUCGAUACAAGAGGCUGUUUUCCUAGCGUGGCUUGCUGCCUUUGGUAAGAAC | | 248 |
| **SMAD4** | WT #1 | GGCUCCGCGGCCGCCCAGGGGGUGGGAGCGGUGAGGGAGCCAGGCGCCCAGC | 144-197 | 54 |
| | G/A-mutant #1 | GGCUCCGCGGCCGCCCAGAGAGUGAGAGCGAGUGAGAGGAGCCAGGCGCCCAGC | | 54 |
| | WT #2 | GGGCAGCGGCGCGGCGCUGAGGAGGGGCGGCCUGGCCGGGACGCCUCGGGGCGGGGGCCGAGGAGCUCUCCGGGCCGCCGGGAAAG CUACGGGCCCGGUGCGUCCG | 254-360 | 107 |
| | G/A-mutant #2 | GGGCAGCGGCGCGGCGCUGAGGAGAGACGGCCUGGCCGAGACGCCUCGAGACGAGAGCCGAGGAGCUCUCCGAGCCGCCGAGGAAAG CUACGAGCCCGGUGCGUCCG | | 107 |

| | | | | |
|---|---|---|---|---|
| **SMAD4** | UTR | AUGCUCAGUGGCUUCUCGACAAGUUGGCAGCAACAACACGGCCCUGGUCGUCGUCGCCGCUGCGGUAACGGAGCGGUUUGGGUGGCG GAGCCUGCGUUCGCGCCUUCCCGCUCUCCUCGGGAGGCCCUUCCUGCUCUCCCCUAGGCUCCGCGGCCGCCCAGGGGGUGGGAGCGG GUGAGGGGAGCCAGGCGCCCAGCGAGAGAGGCCCCCCGCCGCAGGGCGGCCCGGGAGCUCGAGGCGGUCCGGCCCGCGCGGGCAGCG GCGCGGCGCUGAGGAGGGGCGGCCUGGCCGGGACGCCUCGGGGCGGGGGCCGAGGAGCUCUCCGGGCCGCCGGGGAAAGCUACGGGC CCGGUGCGUCCGCGGACCAGCAGCGCGGGAGAGCGGACUCCCCUCGCCACCGCCCGAGCCCAGGUUAUCCUGAAUACAUGUCUAACA AUUUUCCUUGCAACGUUAGCUGUUGUUUUUCACUGUUUCCAAAGGAUCAAAAUUGCUUCAGAAAUUGGAGACAUAUUUGAUUUAAAA GGAAAAACUUGAACAA | | 538 |
| **SMAD7** | WT | **GGG**CGGAGAGCCGCGCAGGGCGCGGGCCGCGCGGGGUGGGGCAGCCGGAGCGCAGGCCCCCGAUCCCCGGCGGGCGCCCCCGGGCCC CCGC | 1-88 (+3) | 91 |
| | G/A-mutant | **GGG**CGGAGAGCCGCGCAGAGCGCGAGCCGCGCGGAGUAGAGCAGCCGGAGCGCAGGCCCCCGAUCCCCGGCGGGCGCCCCCGGGCCC CCGC | | 91 |
| | UTR | CGGAGAGCCGCGCAGGGCGCGGGCCGCGCGGGGUGGGGCAGCCGGAGCGCAGGCCCCCGAUCCCCGGCGGGCGCCCCCGGGCCCCCG CGCGCGCCCCGGCCUCCGGGAGACUGGCGCAUGCCACGGAGCGCCCCUCGGGCCGCCGCCGCUCCUGCCCGGGCCCCUGCUGCUGCU GCUGUCGCCUGCGCCUGCUGCCCCAACUCGGCGCCCGACUUCUUCAUGGUGUGCGGAGGUCAUGUUCGCUCCUUAGCAGGCAAACGA CUUUUCUCCUCGCCUCCUCGCCCCGC | | 287 |
| **SMURF1** | WT | GGACCCCGGCGCCCAGCCCGGAGCCGUAACCUUGAGGCGGCGGCGGCGGGCCGGGCCGGGCCGGCUGGGGGGCGGUGGCGCUGGA UCCGCGGCUGCCC | 206-305 | 100 |
| | G/A-mutant | GGACCCCGGCGCCCAGCCCGGAGCCGUAACCUUGAGACGACGACGACGAAGCCGAGCCGAGCCGAGCUGAGAGACGAUGACGCUGGA UCCGCGGCUGCCC | | 100 |
| | UTR | GGCAGCGGCGGAAGCGGCGAGGGCGGCGGGCGUCCGGCUCUGAGGUGGUGGAGGCGGCGGAGGCGGCGGCGGAGGCGGCGGCGGCUC GGGACUGGGCUCGGCUGGAAGCAGCGAGGGUCAGAGCGCCGCAGCAAGCGCCGAUCUCCCGGCUCGACCAUCCGCCUGCCGCCCGGA CGCCUGGGCCGCGGAGUUUGUGUCCCGGCUCGGACCCCGGCGCCCAGCCCGGAGCCGUAACCUUGAGGCGGCGGCGGCGGGGCCGGG CCGGGCCGGGCUGGGGGGCGGUGGCGCUGGAUCCGCGGCUGCCCGAUCGUUGGCGGGAG | | 320 |
| **TCF7L1** | WT | **GG**GCGCCGGGCCGGGCCGGCAGGGCGCGGCGGCUAGGGGCUCCGAGAGCGGCGGCCCCGGCCCGCGGCCCCACC | 1-74 (+2) | 76 |
| | G/A-mutant | **GG**GCGCCGAGCCGAGCCGAGCAGAGCGCGAGCGGCUAGAGGCUCCGAGAGCGGCGGCCCCGGCCCGCGGCCCCACC | | 76 |
| | UTR | GCGCCGGGCCGGGCCGGGCAGGGCGCGGGCGGCUAGGGGCUCCGAGAGCGGCGGCCCCGGCCCGCGGCCCCACC | | 74 |

**Table S2 Comparison of the prediction methods**

| Predictions tools | G4 predictions | dsRNA predictions | TP | FP | TN | FN | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| cG/cC | 13 | 13 | 10 | 3 | 8 | 5 | 0.66 | 0.72 |
| G4H | 6 | 20 | 5 | 1 | 10 | 10 | 0.33 | 0.90 |
| G4NN | 10 | 16 | 9 | 1 | 10 | 6 | 0.60 | 0.90 |
| RNAfold | 16 | 10 | 10 | 6 | 5 | 5 | 0.66 | 0.45 |
| *In vitro* | + | - | | | | | | |
| Confirmation | 15 | 11 | | | | | | |

TP: True positive, FP: False positive, TN: True negative, FN: False negative

**Table S3 UTRref, RefSeq and Gene-ontology Identification numbers of all candidates**

| PG4 Candidates | | UTRref locus ID | RefSeq transcript | KEGG orthology | KEGG pathway | AmiGO-2, GO class (direct) |
|---|---|---|---|---|---|---|
| **Wnt signaling pathway** | APC | 5HSAR052987 | NM_001127511 | K02085 | ko04310 | |
| | BCL-9L | 5HSAR036862 | NM_182557 | | | canonical Wnt signaling pathway, GO_REF:0000107 |
| | FZD10 as FZD9_10 | 5HSAR051920 | NM_007197 | K02842 | ko04310 | |
| | FZD2 | 5HSAR038614 | NM_001466 | K02235 | ko04310 | |
| | TCF7L1 | 5HSAR039936 | NM_031283 | K04490 | ko04310 | |
| **Apoptosis and Apoptosis multiple species** | AIFM2 | 5HSAR058003 | NM_032797 | | | positive regulation of apoptotic process, PAINT_REF:43735 |
| | APPL1 | 5HSAR053853 | NM_012096 | K08733 | | extrinsic apoptotic signaling pathway in absence of ligand,Reactome:R-HSA-418889 |
| | BAD | 5HSAR038501 | NM_004322 | K02158 | ko04210 , ko04151 | |
| | BAG-1 | 5HSAR059595 | NM_004323 | K09555 | | Apoptotic process, GO_REF:0000037 |
| | BAG-5 | 5HSAR038892 | NM_001015048 | K09559 | | negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway, PMID:24475098 |
| | BCL-2 | 5HSAR038163 | NM_000633 | K02161 | ko04210, ko04151 | |
| | BOK | 5HSAR056739 | NM_032515 | K02561 | ko04215 | |
| | CASP6 | 5HSAR056027 | NM_032992 | K04396 | ko04210 | |
| | CASP8AP2 (as CED-4) | 5HSAR043588 | NM_012115 | K20105 | ko04215 | |
| | CASP9 | 5HSAR042668 | NM_001229 | K04399 | ko04210, ko04215, ko04151 | |
| **TGF-beta signaling pathway** | ACVRIC | 5HSAR034445 | NM_145259 | K13568 | ko04350 | |
| | BMPR1A | 5HSAR049863 | NM_004329 | K04673 | ko04350 | |
| | SMAD2 | 5HSAR048712 | NM_005901 | K04500 | ko04350 | |
| | SMAD4 | 5HSAR030859 | NM_005359 | K04501 | ko04350 | |
| | SMAD7 | 5HSAR056165 | NM_005904 | K19631 | ko04350 | |
| | SMURF1 | 5HSAR043726 | NM_181349 | K04678 | ko04350 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **PI3K-Akt signaling pathway** | PIK3R1 as PIK3R1_2_3 | 5HSAR037471 | NM_181524 | K02649 | ko04151, ko04010, ko4210 | |
| | PIK3R3 as PIK3R1_2_3 | 5HSAR049275 | NM_003629 | K02649 | ko04151, ko04010, ko4210 | |
| | MAP2K1 (MEK1) | 5HSAR052794 | NM_002755 | K04368 | ko04151, ko04010, ko4210 | |
| | MAPK3 (ERK1) as MAPK1_3 | 5HSAR030886 | NM_002746 | K04371 | ko04151 , ko04010, ko4210 | |

**Table S4 Oligonucleotide sequences used for PCR-filling prior to *in vitro* transcription**

| Candidates | | | Name | Sequence 5'-3' |
|---|---|---|---|---|
| **ACVR1C** | WT | #1 | ACR1V_fwd | **TAATACGACTCACTATAGGG**CCGCCCGGCTGCGGGGCCAGTGGCAGGAGCGCCGCGCACCGCCAGCC |
| | | #2 | ACR1V_rev | ACACCCTTTTGAAGTGCGCGGTTGGCTCTAGTCAGTGTGGGCGCCCCCCTCCCCGGCCGCCCCCATCCCACGCCCCCTGCGGCTGGCGGTGCGCGGCGCT |
| | G4mut | #1 | ACR1V_fwd | **TAATACGACTCACTATAGGG**CCGCCCGGCTGCGGGGCCAGTGGCAGGAGCGCCGCGCACCGCCAGCC |
| | | #2 | ACR1V mutG-A rev | ACACCCTTTTGAAGTGCGCGGTTGGCTCTAGTCAGTGTGGGCGCCCCCCTCTCCGGCCGCTCTCATCTCACGCTCTCTGCGGCTGGCGGTGCGCGGCGCT |
| **AIFM2** | WT | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | AIFM2_WT_R | CGCCCGCGCGCTCTCGCTCCGGCTCCCGCTCCCGCTCCCGCTCCCGCTTGGTCGTCTTCCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | AIFM2_G-A-MUT_R | CGCCCGCGCGCTCTCGCTTCGGCTCTCGCTCTCGCTCTCGCTCTCGCTTGGTCGTCTTCCC**TATAGTGAGTCGTATTA** |
| **APC** | WT | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | APC_wt_rev | CCTGAGCCCGCAGGTCCCCCCCGAGCAGCGGCTAGGCTTCCGGCGGCCACACCCCGCCCCCTTGCTACTTGCCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | APC_G/Amut_rev | CCTGAGCCCGCAGGTCTCTCTCGAGCAGCGGCTAGGCTTCCGGCGGCCACACTTCGCTCTCTTGCTACTTGCCC**TATAGTGAGTCGTATTA** |
| **APPL1** | WT | #1 | APPL1wt_fwd | **TAATACGACTCACTATAGG**GCTCGGCGCCTGGAGAAGGCTGTGCGGGCGGGGACGGCTGCAGCCCTTGCCGGAGAGGGCGGGCCGGGGTCAGCTGCGGC |
| | | #2 | APPL1wt_rev | CGTGGCGGAGGGAGGGCGGTGGCAGGAGACGCAGCTGCCGCCCACAGCTCCCCGCGCCGGCCCGCCCGCCGCAGCTGACCCC |
| | G4mut | #1 | APPl1g4mut_fwd | **TAATACGACTCACTATAGG**GCTCGGCGCCTGGAGAAGGCTGTGCGGGCGGGGACGGCTGCAGCCCTTGCCGGAGAGAGCGAGCCGGAGTCAGCTGCGGC |
| | | #2 | APPL1g4mut_rev | CGTGGCGGAGGGAGGGCGGTGGCAGGAGACGCAGCTGCCGCTCACAGCTCCTCGCGCCGGCTCGCTCGCCGCAGCTGACTCC |
| **BAD** | WT | #1 | BAD_WT_fwd | **TAATACGACTCACTATAGGG**TCAGGGGCCTCGAGATCGGGCTTGGGGTGAGACCTGTGCGCCGTCACCACGGGCGGGGCGGGGCCTGGGTCCACCGGGGT |
| | | #2 | BAD_WT_rev | TGGCAGGGAGCTGAGGCTGTCGGCTCAGGACCTCAGTCTCCCCTCAGAACCCCGGTGGACCCAGGCC |

| | | | | |
|---|---|---|---|---|
| | G4mut | #1 | BAD_G-Amut_fwd | **TAATACGACTCACTATAGGG**TCAGGAGCCTCGAGATCGAGCTTGAGGTGAGACCTGTGCGCCGTCACCACGAGCGGAGCGAGGCCTGAGTCCACCGGAGT |
| | | #2 | BAD_G-Amut_rev | TGGCAGGGAGCTGAGGCTGTCGGCTCAGGACCTCAGTCTCCCCTCAGAACTCCGGTGGACTCAGGCC |
| **BAG1** | WT | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | BAG1_WT-rev | TTGTTGACCGCCCAGCGATGGAAGCTGAGCGCGGCGTCTCACAACCCCGCCCGACTACTTCCCAGCCCCGCCCCGGCCTGCCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | BAG1_G-Amut_rev | TTGTTGACCGCCCAGCGATGGAAGCTGAGCGCGGCGTCTCACAACCTCGCTCGACTACTTCTCAGTCTCGTCTCGGCCTGCCC**TATAGTGAGTCGTATTA** |
| **BAG5** | WT | #1 | BAG5_wt_fwd | **TAATACGACTCACTATA**GGCGCGGACGCCGGAGGAGGTGTCCCCGGGTTTAGGGGTGTTCGGCCAGGGGCGGGGCTGCCGGGCCCGGGCGACTGCCGGAG |
| | | #2 | BAG5_rev | CGCGCGATGCGTCGCCGACGCTTCCACGACTTCCGCAGCTCCGGCAGTCGC |
| | G4mut | #1 | BAG5_G-Amut_fwd | **TAATACGACTCACTATA**GGCGCGGACGCCGGAGGAGGTGTCCCCGAGTTTAGGAGTGTTCGGCCAAGAGCAGAGCTGCCGAGCCCGAGCGACTGCCGGAG |
| | | #2 | BAG5_rev | CGCGCGATGCGTCGCCGACGCTTCCACGACTTCCGCAGCTCCGGCAGTCGC |
| **BCL2** | WT | #1 | BCL2_wt_fwd | **TAATACGACTCACTATAGG**GCAAAGCACATCCAATAAAATAGCTGGATTATAACTC |
| | | #2 | BCL2_wt_rev | CCTTCCCAGAGGAAAAGCAACGGGGGCCAACGGCACCTCTCGCCCCAGCTCCCACCCCACGGCCCCCAGAGAAAGAAGAGGAGTTATAATCCAGCTATTT |
| | G4mut | #1 | BCL2_wt_fwd | **TAATACGACTCACTATAGG**GCAAAGCACATCCAATAAAATAGCTGGATTATAACTC |
| | | #2 | BCL2_G4mut_rev | CCTTCCCAGAGGAAAAGCAACGGGGGCCAACGGCACCTCTCGCTCCAGCTCTCATCTCACGGCTCTCAGAGAAAGAAGAGGAGTTATAATCCAGCTATTT |
| **BCL9L** | WT | #1 | BCL9L_wt_fwd | **TAATACGACTCACTATAGGG**CGCTCGCTCGCTCTGCCTCTCCGCCCGGGCTCTGCCGAAGGGGGCGGGGTGGGGGTGCAGGGCGGGGGGAGGGGAGGCTC |
| | | #2 | BCL9L_wt_rev | CCAGTAACGCTGGCTCGGATTCCTCCCCGACCGCAAGAATGCAGGAGCCTCCCCTCCCCCCGCCC |
| | G4mut | #1 | BCL9L_G-Amut_fwd | **TAATACGACTCACTATAGGG**CGCTCGCTCGCTCTGCCTCTCCGCCCGAGCTCTGCCGAAGAGAGCGAGATGAGAGTGCAGAGCAGAGAGAGAGAAGGCTC |
| | | #2 | BCL9L_G-Amut_rev | CCAGTAACGCTGGCTCGGATTCCTCCTCGACCGCAAGAATGCAGGAGCCTTCTCTCTCTCTGCTC |
| **BMPR1A** | WT | #1 | T72G | **TAATACGACTCACTATA**GG |
| | | #2 | BMPR1A_WT_R | GCCTAAGCCTTCCTGCCCGCCCTTGGCCCTCCTCCCTCCTCGCTCCTGCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T72G | **TAATACGACTCACTATA**GG |

| | | | | |
|---|---|---|---|---|
| | | #2 | BMPR1A_GA_Mu_R | GCCTAAGCTTTCCTGCTCGCTCTTGGCTCTCTTCTCTCTTCGCTCCTGCC**TATAGTGAGTCGTATTA** |
| **BOK** | WT | #1 | BOK_WT_fwd | **TAATACGACTCACTATAGGG**AAGAGCGCGGGAAGCCCCGTGGACCTGGCGCTCCCGGCTCGGGCGTGGACGGGGCGGGCGCCGGGGCGGGGCGCGCGTCC |
| | | #2 | BOK_WT-rev | GCCGCCGACGACGACCTCGACCCTTCCATTCAGACCCGCGAGGACGCGCGCC |
| | G4mut | #1 | BOK_G-Amut_fwd | **TAATACGACTCACTATAGGG**AAGAGCGCGGGAAGCCCCGTGGACCTGGCGCTCCCGGCTCGGGCGTGGACGAGACGAGCGCCGAGACGAGGCGCGCGTCC |
| | | #2 | BOK_WT-rev | GCCGCCGACGACGACCTCGACCCTTCCATTCAGACCCGCGAGGACGCGCGCC |
| **CASP6** | WT | #1 | T72G | **TAATACGACTCACTATA**GG |
| | | #2 | Casp6-5UTR-Trx wt | TGCAGCCAAACGCGCAGCCAGACACCTTGCCCTCCTCTTCCTGAAGCCACAGGCTCCCGGGCCCGGCCCCGCCCTCGGCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T72G | **TAATACGACTCACTATA**GG |
| | | #2 | Casp6-5UTR-Trx mut | TGCAGCCAAACGCGCAGCCAGACACCTTGCCCTCCTCTTCCTGAAGCCACAGGCTCCCGGGCCCGGCTTCGCTCTCGGCC**TATAGTGAGTCGTATTA** |
| **CASP8AP2** | WT | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | CASP8AP2_WT_rev | CTTCCTTAGCAGTCCCTCACGGACCGGGCCCCTACGACAACTTACCCGCCCTGCCCGGCCCAAGACAACCCGGTTCCTTTCCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | CASP8AP2_G-Amut_rev | CTTCCTTAGCAGTCCCTCACGGACCGGGCCTCTACGACAACTTACTCGCTCTGCTCGGCTCAAGACAACTCGGTTCCTTTCCC**TATAGTGAGTCGTATTA** |
| **CASP9** | WT | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | CASP9_WT_rev | GACTCCAGGCCGCCTCAGTCCGCTTCCGGGCCTCGGCCGCCCGCCCCAGTCCCCAGGACCCGCCCCCGCCCCAGGGCCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | CASP9_G-Amut_rev | GACTCCAGGCCGCCTCAGTCCGCTTCCGGGCCTCGGCCGCTCGCCTCAGTCCTCAGGACTCGCTCTCGCTCCAGGGCCC**TATAGTGAGTCGTATTA** |
| **FZD10** | WT | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | FZD10_WT_rev | CCACCCCTGCTCGGGCCCCCGCCCATGCCCGCCCGGCCTGGCCCGAGCGCTGCGCACAGCGCCCCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | FZD10_G-A-mut_rev | CCACCCCTGCTCGGGCTCTCGCTCATGCTCGCTCGGCCTGGCTCGAGCGCTGCGCACAGCGCCCC**TATAGTGAGTCGTATTA** |

| | | | | |
|---|---|---|---|---|
| **FZD2** | WT | #1 | FZD2_WT_fwd | **TAATACGACTCACTATA**GGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGTCTCC |
| | | #2 | FZD2_WT_rev | AGTCATGGTGGCTAGCGCTGGCCGCCGCCCCCCACCCGGCTCCTTGGCGCCCCCCCCGCCCCCGCCCCCAACCCGGAGACTGCGCTTCTTCCCC |
| | G4mut | #1 | FZD2_WT_fwd | **TAATACGACTCACTATA**GGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGTCTCC |
| | | #2 | FZD2_G-A_fwd | AGTCATGGTGGCTAGCGCTGGCCGCCGCCCCCCACCCGGCTCCTTGGCGCTTCCTTCGCTCTCGCTTCCAACCCGGAGACTGCGCTTCTTCCCC |
| **MAP2K1** | WT | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | MAP2K1_WT_R | GGCCGGTAGCGGTCTCAGTGGACCCCCGCCCCACCCGCCCGGGACTCGGCTTCGCGCGCAGAGAGCCGAAAGGCTGCCC**TATAGTGAGTCGTATTA** |
| | G4mut | #1 | T73G | **TAATACGACTCACTATA**GGG |
| | | #2 | MAP2K1_G/A-MUT_R | GGCCGGTAGCGGTCTCAGTGGACTCTCGTCTCACTCGCTCGGGACTCGGCTTCGCGCGCAGAGAGCCGAAAGGCTGCCC**TATAGTGAGTCGTATTA** |
| **MAPK3** | WT | #1 | MAPK3 fwd | **TAATACGACTCACTATA**GGGCGGGTGACAGGCAGGCGGGAAGGGGCG |
| | | #2 | MAPK3 rev | CTCCACTCCTCCCCTCCCACCGCCCTCCTCCCCACGGCGGCCCCGCCCGAGGCCCCGCCCCTTCCCGCCTGCCTG |
| | G4mut | #1 | MAPK3 fwd | **TAATACGACTCACTATA**GGGCGGGTGACAGGCAGGCGGGAAGGGGCG |
| | | #2 | MAPK3 mut rev | CTCCACTCCTCTCCTCTCACCGCTCTCCTCCTCACGGCGGCCCCGCCCGAGGCCCCGCCCCTTCCCGCCTGCCTG |
| | 1er G/Amut | #1 | MAPK3 1erG>Afwd | **TAATACGACTCACTATA**GGGCGGGTGACAGGCAGGCGAGAAGAGACG |
| | | #2 | MAPK3 1erG>Arev | CTCCACTCCTCCCCTCCCACCGCCCTCCTCCCCACGGCGGCCTCGCTCGAGGCTCCGTCTCTTCTCGCCTGCCTG |
| | G/Amut All | #1 | MAPK3 1erG>Afwd | **TAATACGACTCACTATA**GGGCGGGTGACAGGCAGGCGAGAAGAGACG |
| | | #2 | MAPK3_allG>Arev | CTCCACTCCTCTCCTCTCACCGCTCTCCTCCTCACGGCGGCCTCGCTCGAGGCTCCGTCTCTTCTCGCCTGCCTG |
| **PIK3R1** | WT | #1 | T72G | **TAATACGACTCACTATA**GG |
| | | #2 | PIK3R1_WT_rev | CTCGGCTCGGCTCGTCCCTCCGCCCCACGCCCCGCCCGGCAGTGCGTCCGGGCCACCGCCTATAGTGAGTCGTATTA |
| | G4mut | #1 | T72G | **TAATACGACTCACTATA**GG |
| | | #2 | PIK3R1_G4mut_rev | CTCGGCTCGGCTCGTCTCTCCGCTTCACGCTTCGCTCGGCAGTGCGTTCGGGTCATCGCCTATAGTGAGTCGTATTA |
| **PIK3R3** | WT | #1 | PIK3R3_WT_fwd | **TAATACGACTCACTATAGG**CCTCAGCGGGTGGGCAGCATGGGGCGGGGAGGGTGTCCCCTCCGCGCCGTTAAAATG |

| | | | | |
|---|---|---|---|---|
| | | #2 | PIK3R3_rev | TGCCCGGACTCCAGCCACTAGAGTTTCATTTTAACGGCGCGGAGGGG |
| | G4mut | #1 | PIK3R3_G-Amut_fwd | **TAATACGACTCACTATAGG**CCTCAGCGAGTGAGCAGCATGAGGCGAAGAGAGTGTCCCCTCCGCGCCGTTAAAATG |
| | | #2 | PIK3R3_rev | TGCCCGGACTCCAGCCACTAGAGTTTCATTTTAACGGCGCGGAGGGG |
| **SMAD2** | WT | #1 | SMAD2wt_fwd | **TAATACGACTCACTATAGG**GCGCCCGGGCCGCCGGCCGGGCCC |
| | | #2 | SMAD2wt_rev | GGGAATGGGCGATTGGAGGCGGAACTGAAAACACTCCCGGCCGCCGTCTTCCCGCCCCGCCCCCAGGCCCGGGCCCGGCCGGCGGCCCGG |
| | G4mut | #1 | SMAD2wt_fwd | **TAATACGACTCACTATAGG**GCGCCCGGGCCGCCGGCCGGGCCC |
| | | #2 | SMAD2_G-A_rev | GGGAATGGGCGATTGGAGGCGGAACTGAAAACACTCCCGGCCGCCGTCTTCTCGCTCCGCTCTCAGGCTCGGGCTCGGCCGGCGGCCCGG |
| **SMAD4** | WT | #1 | T72G | **TAATACGACTCACTATA**GG |
| | | #2 | SMAD4wt_rev | GCTGGGCGCCTGGCTCCCCTCACCCGCTCCCACCCCCTGGGCGGCCGCGGAGCCTATAGTGAGTCGTATTA |
| | G4mut | #1 | T72G | **TAATACGACTCACTATA**GG |
| | | #2 | SMAD4g4mut_rev | GCTGGGCGCCTGGCTCCTCTCACTCGCTCTCACTCTCTGGGCGGCCGCGGAGCCTATAGTGAGTCGTATTA |
| **SMAD4 #2** | WT | #1 | SMAD4_2_WT_fwd | **TAATACGACTCACTATAGGG**CAGCGGCGCGGCGCTGAGGAGGGGCGGCCTGGCCGGGACGCCTCGGGGCGGGGGCCGAGGAGCTCTCC |
| | | #2 | SMAD4_2_WT_rev | CGGACGCACCGGGCCCGTAGCTTTCCCCGGCGGCCCGGAGAGCTCCTCGG |
| | G4mut | #1 | SMAD4_2_G4mut_fwd | **TAATACGACTCACTATAGGG**CAGCGGCGCGGCGCTGAGGAGAGACGGCCTGGCCGAGACGCCTCGAGACGAGAGCCGAGGAGCTCTCC |
| | | #2 | SMAD4_2_G4mut_rev | CGGACGCACCGGGCTCGTAGCTTTCCTCGGCGGCTCGGAGAGCTCCTCGG |
| **SMAD7** | WT | #1 | SMAD7wt_fwd | **TAATACGACTCACTATAGGG**CGGAGAGCCGCGCAGGGCGCGGGCCGCGCGGGGTGGGGCAGCCGGAGCGCAGGCCCC |
| | | #2 | SMAD7wt_rev | GCGGGGGCCCGGGGGCGCCCGCCGGGGATCGGGGGCCTGCGCTCCGGCTGC |
| | G4mut | #1 | SMAD7G/A_fwd | **TAATACGACTCACTATAGGG**CGGAGAGCCGCGCAGAGCGCGAGCCGCGCGGAGTAGAGCAGCCGGAGCGCAGGCCCC |
| | | #2 | SMAD7wt_rev | GCGGGGGCCCGGGGGCGCCCGCCGGGGATCGGGGGCCTGCGCTCCGGCTGC |
| **SMURF1** | WT | #1 | SMURF1_Fwd | **TAATACGACTCACTATA**GGACCCCGGCGCCCAGCCCGGAGCCGTAACCTTGAG |
| | | #2 | SMURF1_WT_rev | GGGCAGCCGCGGATCCAGCGCCACCGCCCCCCAGCCCGGCCCGGCCCGGCCCCGCCGCCGCCGCCTCAAGGTTACGGCTCCGGG |

| | | | | |
|---|---|---|---|---|
| | G4mut | #1 | SMURF1_Fwd | **TAATACGACTCACTATA**GGACCCCGGCGCCCAGCCCGGAGCCGTAACCTTGAG |
| | | #2 | SMURF1_G4_mut_rev | GGGCAGCCGCGGATCCAGCGTCATCGTCTCTCAGCTCGGCTCGGCTCGGCTTCGTCGTCGTCGTCTCAAGGTTACGGCTCCGGG |
| **TCF7L1** | WT | #1 | TCF7L1_WT_fwd | **TAATACGACTCACTATA**GGGCGCCGGGCCGGGCCGGGCAGGGCGCGGGCGGCTAGGGGCTCCGAGAGCGGCGGCCC |
| | | #2 | TCF7L1_WT_rev | GGTGGGGCCGCGGGCCGGGGCCGCCGCTCTCGGAGCC |
| | G4mut | #1 | TCF7L1_G-A_fwd | **TAATACGACTCACTATA**GGGCGCCGAGCCGAGCCGAGCAGAGCGCGAGCGGCTAGAGGCTCCGAGAGCGGCGGCCC |
| | | #2 | TCF7L1_WT_rev | GGTGGGGCCGCGGGCCGGGGCCGCCGCTCTCGGAGCC |

**Table S5 Oligonucleotide sequences used for PCR-filling prior to cloning**

| Construct | | | | Sequences 5'-3' |
|---|---|---|---|---|
| APC WT | PCR1 | #1 | APC fwd WT_1 | CCTCCCACAAGATGGCGGAGGGCAAGTAGCAAGGGGGCGGGGTGTGGCCGCCGGAAGCCTAGCCGCTGCTCGGGGGGGACCTGCGGGCTCAGGCCCGGG |
| | | #2 | APC rev_1 | CTGAGTGCTTCACCTTCCTCACCAACAGCCAACAACAGTACCTGGGAACAGCATCGAGCCAACCTCGGTCCGCAGCTCCCGGGCCTGAGCCCGCAGGTC |
| | PCR2 | #1 | Nhe1_APC fwd_2 | AGAGCTAGCAGTCTTCCCACCTCCCACAAGATGGCGGAGGG |
| | | #2 | Nhe1_APC rev_2 | AGTCAGTGCTAGCAGGGGGCGCCGAGGCCCGAGAAGGCAACTGAGTGCTTCACCTTCC |
| | | #3 | | 1 µL of PCR1 product |
| APC G/A-mut | PCR1 | #1 | APC fwd G/A-mut_1 | CCTCCCACAAGATGGCGGAGGGCAAGTAGCAAGAGAGCGAAGTGTGGCCGCCGGAAGCCTAGCCGCTGCTCGAGAGAGACCTGCGGGCTCAGGCCCGGG |
| | | #2 | APC rev_1 | CTGAGTGCTTCACCTTCCTCACCAACAGCCAACAACAGTACCTGGGAACAGCATCGAGCCAACCTCGGTCCGCAGCTCCCGGGCCTGAGCCCGCAGGTC |
| | PCR2 | #1 | Nhe1_APC fwd_2 | AGAGCTAGCAGTCTTCCCACCTCCCACAAGATGGCGGAGGG |
| | | #2 | Nhe1_APC rev_2 | AGTCAGTGCTAGCAGGGGGCGCCGAGGCCCGAGAAGGCAACTGAGTGCTTCACCTTCC |
| | | #3 | | 1 µL of PCR1 product |
| BAG-1 WT | PCR1 | #1 | Nhe1_BAG1_wt_fwd | AGAGCTAGCAGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTC |
| | | #2 | Nhe1_BAG1_rev | AGTCAGTGCTAGCGCCCGCACTTGTTGACCGCCCAGCGATGGAAGCTGAGCGCGGCGTCTCACAACC |
| BAG-1 G/A-mut | PCR1 | #1 | Nhe1_BAG1_G-Amut_fwd | AGAGCTAGCAGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTC |
| | | #2 | Nhe1_BAG1_rev | AGTCAGTGCTAGCGCCCGCACTTGTTGACCGCCCAGCGATGGAAGCTGAGCGCGGCGTCTCACAACC |
| CASP8AP2 WT | PCR1 | #1 | CASP8AP2_wt_fwd | CCCGGTCCGTGAGGGACTGCTAAGGAAGAGGCTGCATGGCGCGGTAGTCCCCCGAGTGGAGGTCGGCTGCCCCTGGGAAACCAGAGAGTCGGAGGGAGTC |
| | | #2 | CASP8AP2_wt_rev | TCCTTTTGAGAGTCACCACCTGCGGCCAGACGTCAGGCGCGGCCCTTCCCCGACCTACTTGGCCGCTCCAGATGGACTCCCTCCGACTCTCTGGTTTC |
| | PCR2 | #1 | Nhe1_CASP8AP2_wt_fwd | AGAGCTAGCAAAGGAACCGGGTTGTCTTGGGCCGGGCAGGGCGGGTAAGTTGTCGTAGGGGCCCGGTCCGTGAGGGACTGC |

| | | | | |
|---|---|---|---|---|
| | | #2 | Nhe1_CASP8AP2_wt_rev | AGTCAGTGCTAGCGATCCTATTTCCTTTTGAGAGTCACCACC |
| | | #3 | | 1 µL of PCR1 product |
| CASP8AP2 G/A-mut | PCR1 | #1 | CASP8AP2_wt_fwd | CCCGGTCCGTGAGGGACTGCTAAGGAAGAGGCTGCATGGCGCGGTAGTCCCCCGAGTGGAGGTCGGCTGCCCCTGGGAAACCAGAGAGTCGGAGGGAGTC |
| | | #2 | CASP8AP2_wt_rev | TCCTTTTGAGAGTCACCACCTGCGGCCAGACGTCAGGCGCGGCCCTTCCCCGACCTACTTGGCCGCTCCAGATGGACTCCCTCCGACTCTCTGGTTTC |
| | PCR2 | #1 | Nhe1_CASP8AP2_G-Amut_fwd | AGAGCTAGCAAAGGAACCGAGTTGTCTTGAGCCGAGCAGAGCGAGTAAGTTGTCGTAGAGGCCCGGTCCGTGAGGGACTGC |
| | | #2 | Nhe1_CASP8AP2_wt_rev | AGTCAGTGCTAGCGATCCTATTTCCTTTTGAGAGTCACCACC |
| | | #3 | | 1 µL of PCR1 product |
| MAPK3 WT | PCR1 | #1 | MAPK3-1 | AGAGCTAGCCTGGCGCGCGCGGCCCTGCGGGTGACAGGCAGGCG |
| | | #2 | MAPK3-2wt | AGTCAGTGCTAGCCTCCACTCCTCCCCTCCCACCGCCCTCCTCCCCACGGCGGCCCCGCCCGAGGCCCCGCCCCTTCCCGCCTGCCTGTCACCCGC |
| MAPK3 1st G/A-mut | PCR1 | #1 | MAPK3-1 | AGAGCTAGCCTGGCGCGCGCGGCCCTGCGGGTGACAGGCAGGCG |
| | | #2 | MAPK3-2mut1 | AGTCAGTGCTAGCCTCCACTCCTCCCCTCCCACCGCCCTCCTCCCCACGGCGGCCTCGCTCGAGGCTCCGTCTCTTCTCGCCTGCCTGTCACCCGC |
| MAKP3 G/A-mut (2nd) | PCR1 | #1 | MAPK3-1 | AGAGCTAGCCTGGCGCGCGCGGCCCTGCGGGTGACAGGCAGGCG |
| | | #2 | MAPK3-2mut2 | AGTCAGTGCTAGCCTCCACTCCTCTCCTCTCACCGCTCTCCTCCTCACGGCGGCCCCGCCCGAGGCCCCGCCCCTTCCCGCCTGCCTGTCACCCGC |
| MAPK3 double G/A-mut | PCR1 | #1 | MAPK3-1 | AGAGCTAGCCTGGCGCGCGCGGCCCTGCGGGTGACAGGCAGGCG |
| | | #2 | MAPK3-2mut1_2 | AGTCAGTGCTAGCCTCCACTCCTCTCCTCTCACCGCTCTCCTCCTCACGGCGGCCTCGCTCGAGGCTCCGTCTCTTCTCGCCTGCCTGTCACCCGC |

## Supplementary figures

### Figure S1

Available online at URL

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0208363#sec025

### Figure S2

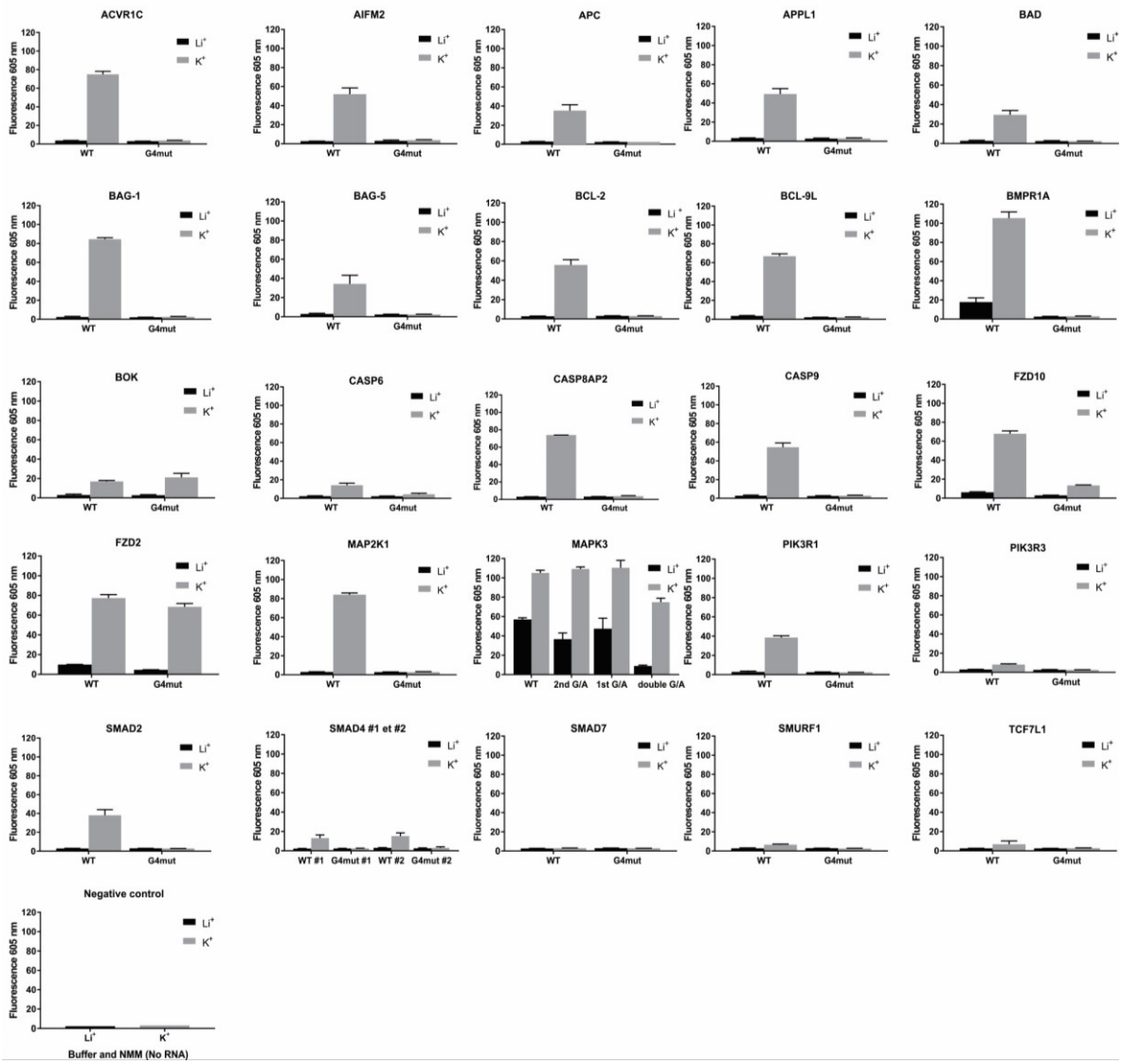



**Figure S2** NMM assay of all candidates.

The fluorescence emission peaks at 605 nm under the different conditions: Black Li+, Gray K+. Each bar represents the mean of 3 independent experiments, the error bars represent the standard deviations.

**Figure S3**

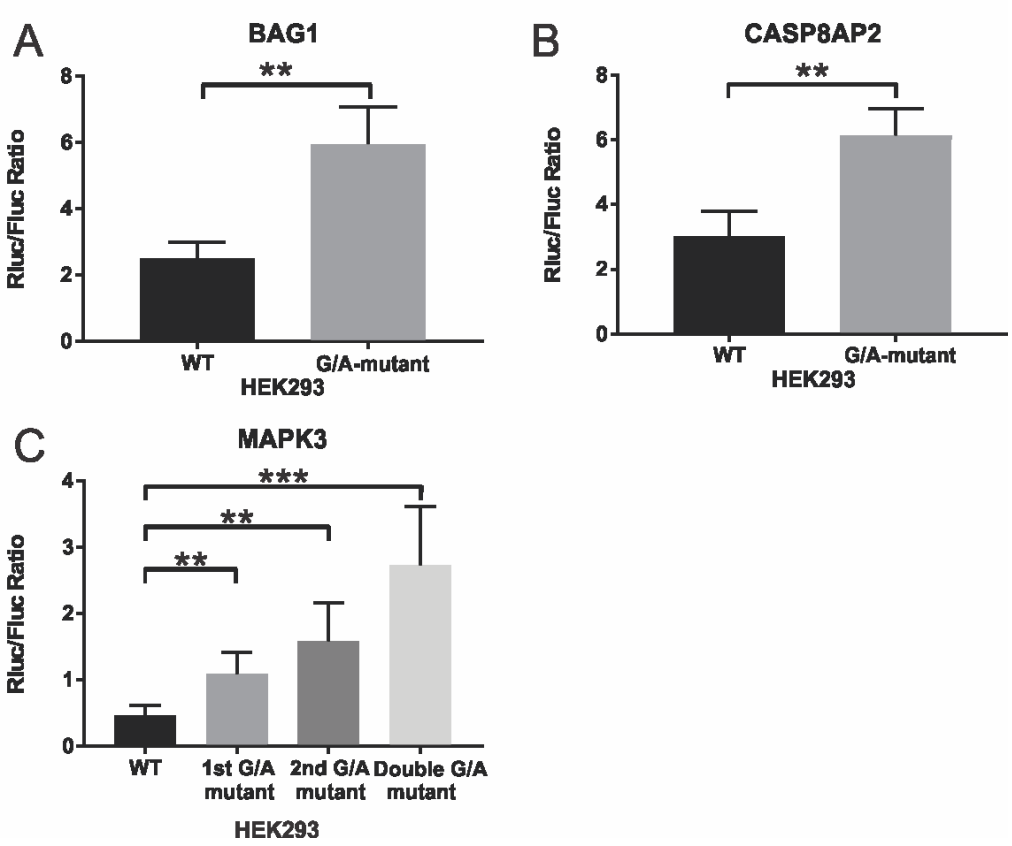



**Figure S3** In cellulo luciferase assay in HEK293 cells.

Results for A) BAG-1 and C) CASP8AP2 from the Apoptosis set; and C) MAPK3 from the PI3-K set. Results are shown as means of the Rluc expression normalized over the Fluc transfection control. The WT results are in black and the G/A-mutants are in different shades of gray. The error bars represent the standard deviations. Statistical difference was measured using an unpaired Student t-test with a n=3 for BAG-1 and CASP8AP2 and n=5 for MAPK3. *P-value < 0.05 **P-value < 0.01 ***P-value < 0.001

**Figure S4**

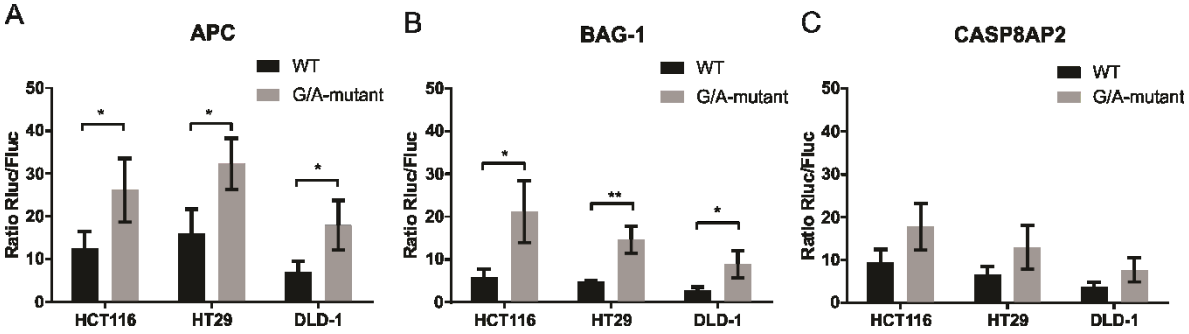



**Figure S4** In cellulo luciferase assay in colorectal cancer cell lines.

The WT and the G/A-mutant full-length 5'UTRs were inserted upstream of the Renilla luciferase (Rluc) reporter gene and used for transfection. The G mutated to A were the same as those in the in vitro assays. A) APC, B) BAG-1 and C) CASP8AP2. The results are shown as the means of the Rluc expression normalized over the Fluc transfection control in the three colorectal cell lines HCT116, HT29 and DLD-1. The WT results are in black and the G/A-mutants' results are in gray. The error bars represent the standard deviations. Statistical difference was measured using an unpaired Student t-test with a n=3 *P-value < 0.05 **P-value < 0.01

**ANNEXE 5 Supplementary data Article 5**

# Supplementary data

## Article 5 – G-quadruplex located in the 5'UTR of the BAG-1 mRNA affects both its cap-dependent and cap-independent translation through global secondary structure maintenance

### Supplementary Material and Methods

BAG-1 endogenous RNA levels in CRC cell lines
Western blot of endogenous BAG-1 in CRC cell lysates

### Supplementary Figures and Legends

**Supplementary Figure S1.** BAG-1 protein isoforms' expression levels in the supplementary paired tissues samples of colorectal tumors at different stages and their adjacent healthy tissue (margin).
**Supplementary Figure S2.** BAG-1 mRNA and protein isoforms' expression levels in normal intestinal epithelial and CRC cell lines.
**Supplementary Figure S3.** RNA and protein isoforms' expression levels of the reporter assays of the complete 5'UTR of BAG-1 with both the mutated rG4 and the mutated 1L or 1M start codons.
**Supplementary Figure S4.** The BAG-1 5'UTR possesses a repressive uORF located at position 254.
**Supplementary Figure S5.** Control of the bicistronic constructions' integrity.
**Supplementary Figure S6.** Secondary structure elucidated by SHAPE of the minimal IRES region of the BAG-1 5'UTR.
**Supplementary Figure S7.** Comparison of the IRES secondary structure elucidated by SHAPE using the WT complete BAG-1 5'UTR sequence with the structure elucidated by Pickering *et al.* 2004.

### Supplementary Tables

**Supplementary Table S1.** Clinicopathological parameters of the CRC patients.
**Supplementary Table S2.** Translation initiation efficiency of the start codons of the BAG-1 5'UTR.
**Supplementary Table S3**. List of primers and oligonucleotides used in this study.
**Supplementary Table S4.** Sequences of all of the 5'UTRs tested, those of the rG4 and the start codons that were mutated, the SHAPE WT and the mutated sequences.
**Supplementary Table S5.** Sequences of the transfected mono- and bicistronic mRNAs.

## Supplementary Material and Methods

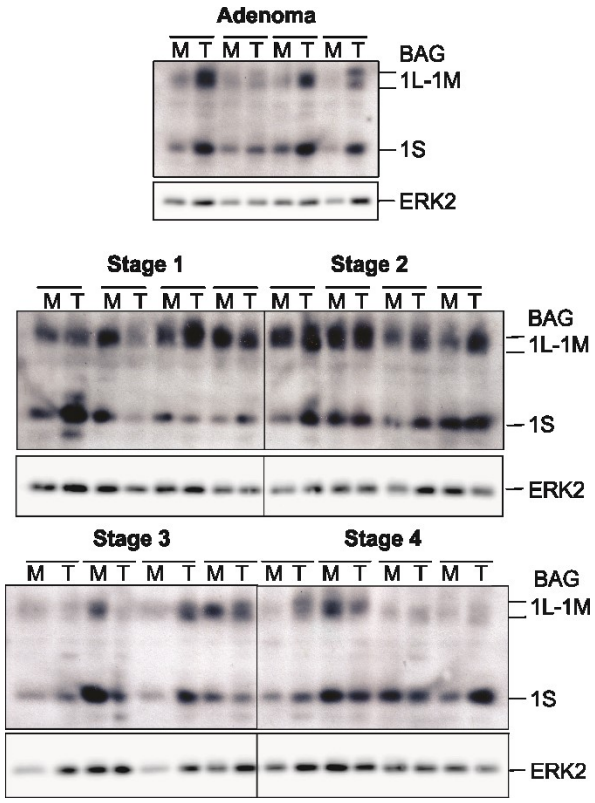
### BAG-1 endogenous RNA levels in CRC cell lines

The cDNAs resulting from the reverse-transcription (RT) of the total RNA extracted from diverse normal and cancerous colorectal cell lines were obtained from the J. Carrier biobank. The qPCR reactions were performed by the RNomics Platform as described in the main manuscript.

### Western blot of endogenous BAG-1 in CRC cell lysates

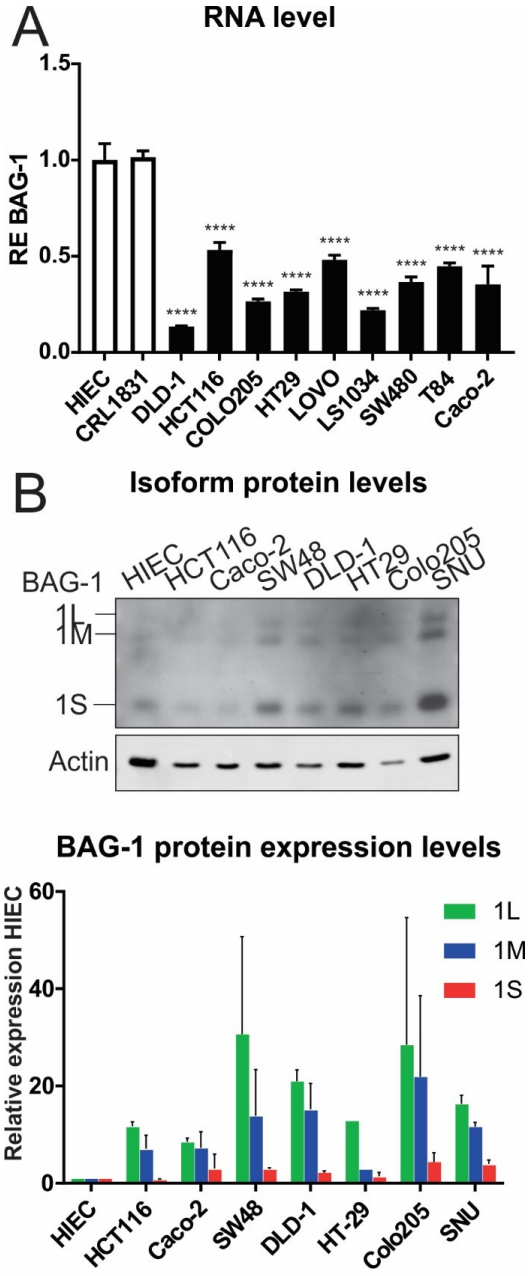
Pooled protein lysates (10 µg) of the colorectal cell lines (HIEC, HCT116, CACO-2/15, SW48, DLD-1, HT-29, Colo205 and SNU) cultured under serum starvation condition were loaded on a 10 % SDS-PAGE gel which was migrated for 2 h 15 min at 150 V and transferred for 1 h at 100 V on a polyvinylidene difluoride (PVDF) membrane which was then blocked 15 min at room temperature in phoshate buffered saline (PBS) with 4% (w/v) nonfat drymilk (PBS-milk 4 %). The Western blot used to detect the BAG-1 protein isoforms was performed as described in the main manuscript, with the exception that β-actin was used as the loading control. After stripping of the membrane in 0.5 N NaOH twice for 10 min, and a thorough washing in PBS, the membrane was blocked in PBS-milk 4 % for 15 min and then incubated for 1 h at room temperature with the primary antibody, mouse anti-β-actin (AS441, Sigma), diluted 1: 1000 in the blocking buffer. After washes in PBS-T, the membrane was incubated for 1 h at room temperature with the secondary antibody, anti-mouse IgG (H+L) (IRDye 800CW, Li-Cor), diluted 1:10 000 in PBS-Milk 4 %. After 3 washes with PBS-T, the membrane was revealed using the Li-Cor Odyssey system.

## Supplementary Figures and Legends



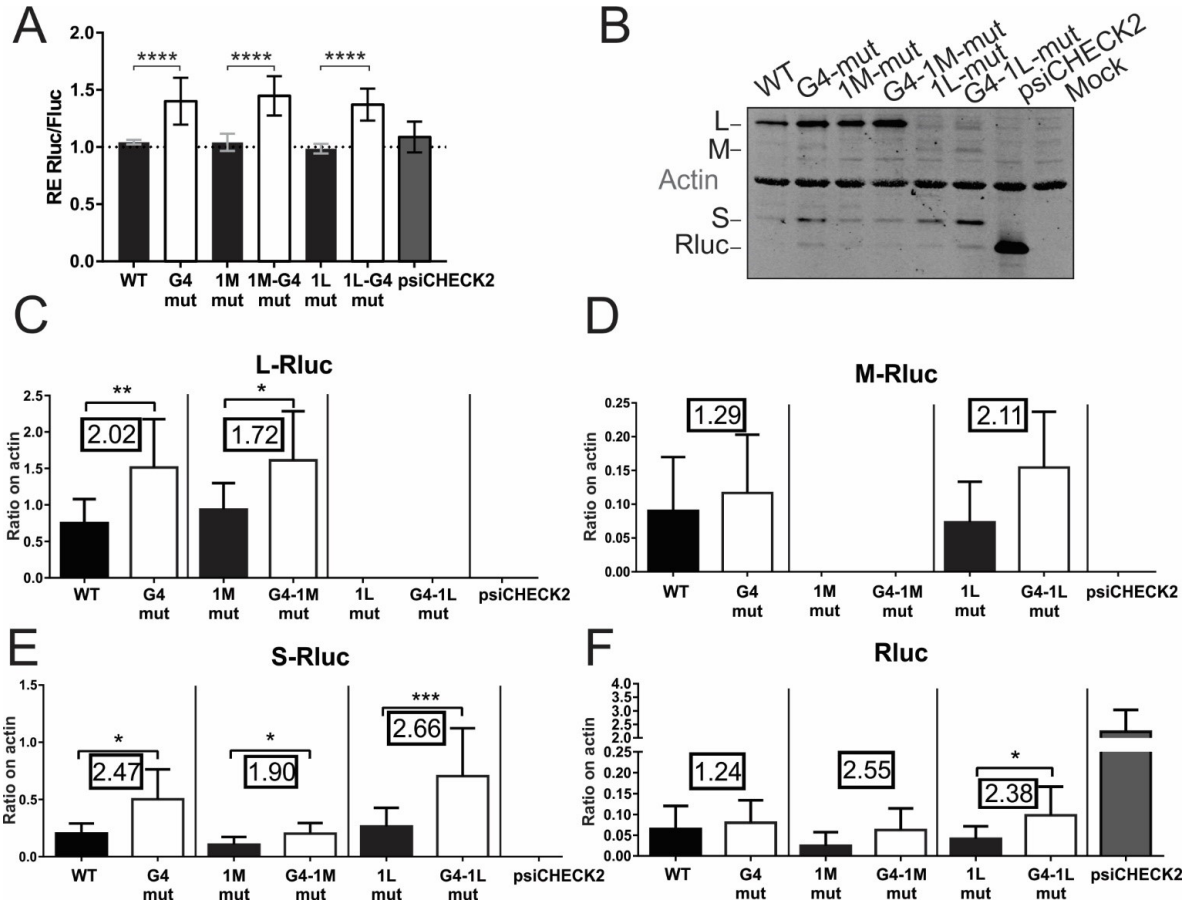

**Supplementary Figure S1**. BAG-1 protein isoforms' expression levels in the supplementary paired tissues samples of colorectal tumors at different stages and their adjacent healthy tissue (margin).

Protein expression levels, as measured by Western blot, of the three BAG-1 isoforms in the same pairs of margin (M)-tumor tissues (T) as in (**Figure 2A**). ERK2 is used as the loading control (*n*=4 for all stages).



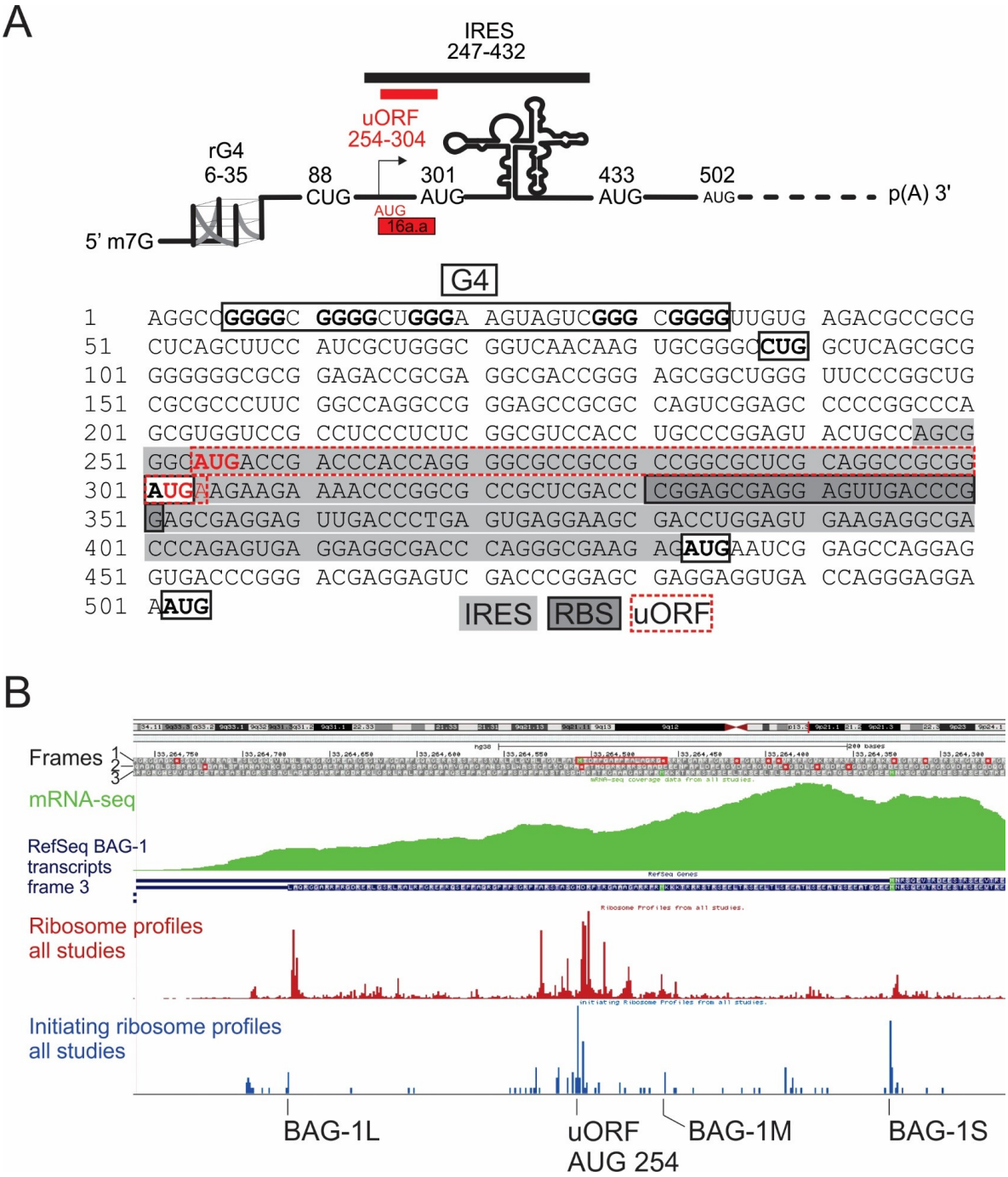

**Supplementary Figure S2.** BAG-1 mRNA and protein isoforms' expression levels in normal intestinal epithelial and CRC cell lines.

(A) Relative RNA levels (RE) of BAG-1 as measured by RT-qPCR. The normal colorectal epithelial cell lines are in white, while the CRC cell lines ares in black. The bars represent the means with their standard deviations ($n$=3). The statistical test performed is a one-way ANOVA with Dunnett's Multiple comparison test. Statistical difference of the BAG-1 RNA RE levels of all cell lines were compared to the RE level of HIEC cells normalised at 1 ****$P \leq 0.0001$. (B) *Top:* Representative immunoblot of the BAG-1 protein isoforms in the normal HIEC cell line and in seven CRC cell lines. β-actin was used as the loading control. *Bottom:* Relative expression levels of the three protein isoforms compare to the HIEC normal cell line. The relative expression was measured as the isoforms' band densities normalised with the actin loading control, all relative to the HIEC band density which was set to 1. The bars are the means and standard deviation ($n$=3).



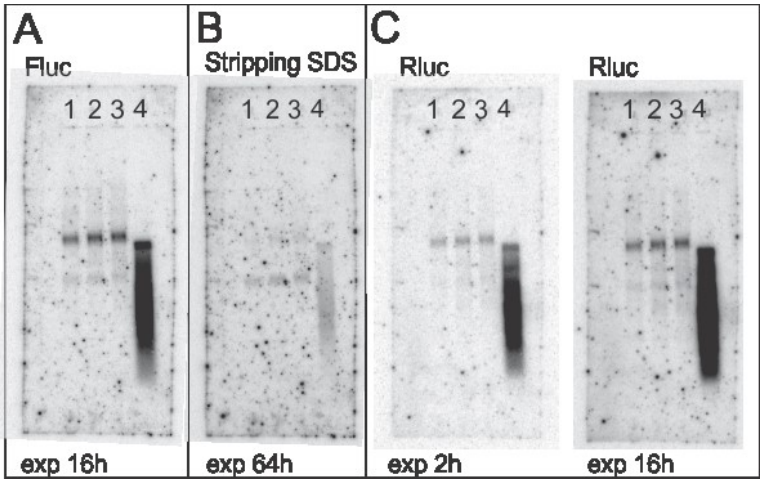

**Supplementary Figure S3.** RNA and protein isoforms' expression levels of the reporter assays of the complete 5'UTR of BAG-1 with both the mutated rG4 and the mutated 1L or 1M start codons.

(A) The relative expression levels of the Rluc RNA, normalised over the Fluc RNA after the transfections of the different mutated constructions, as measured by RT-qPCR. The bar labeled psiCHECK-2 represents the reporter plasmid without the BAG-1 5'UTR insertion. The statistical test performed is a two-way ANOVA with Tukey's multiple comparison test, ($n$=3), ****$P$≤0.0001. (**B**) Representative immunoblot of the Rluc N-extension protein isoforms' expression levels of the rG4 and either the 1L or the 1M start codon mutations. The psiCHECK-2 transfection lane represents the canonical Rluc without the N-terminal extension. Mock represents the untransfected control. β-actin was used as the loading control. (C-F) Quantification of the protein level of each isoform, normalised over the β-actin loading control. (C) L-Rluc, (D) M-Rluc, (E) S-Rluc, (F) Rluc. The boxed value is the fold-change in the protein level of the rG4mut construction over that of the WT. The statistical test performed is a Mann-Whitney test, ($n$=3), *$P$≤0.05, **$P$≤0.001, *** $P$≤0.0005.



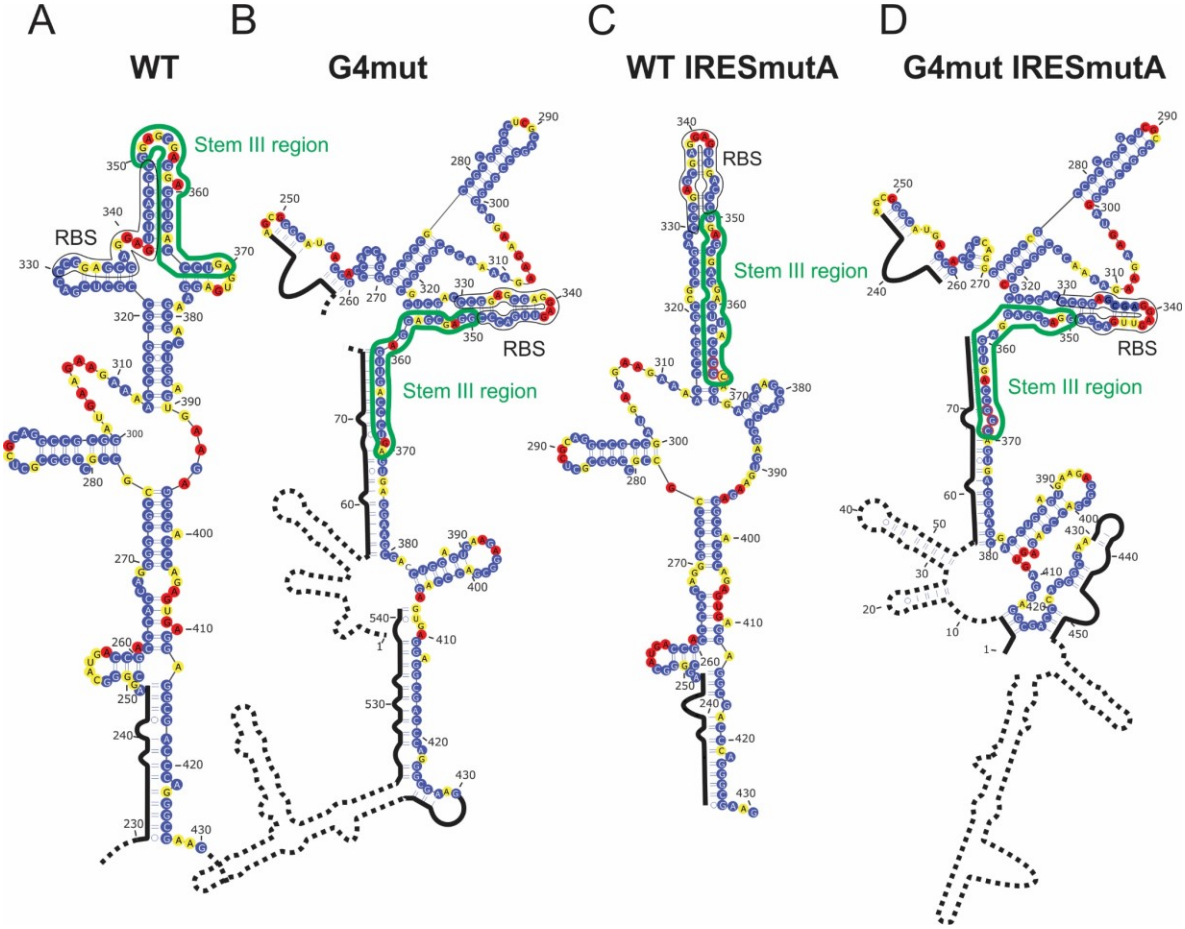

**Supplementary Figure S4**. The BAG-1 5'UTR possesses a repressive uORF located at position 254.

(A) Scheme of the BAG-1 5'UTR organisation showing the position of the possible uORF that is located within the 5'UTR. (B) Genome-browser view of the aggregate of the multiple ribosome-profiling studies demonstrating the initiation of translation at the AUG position 254.



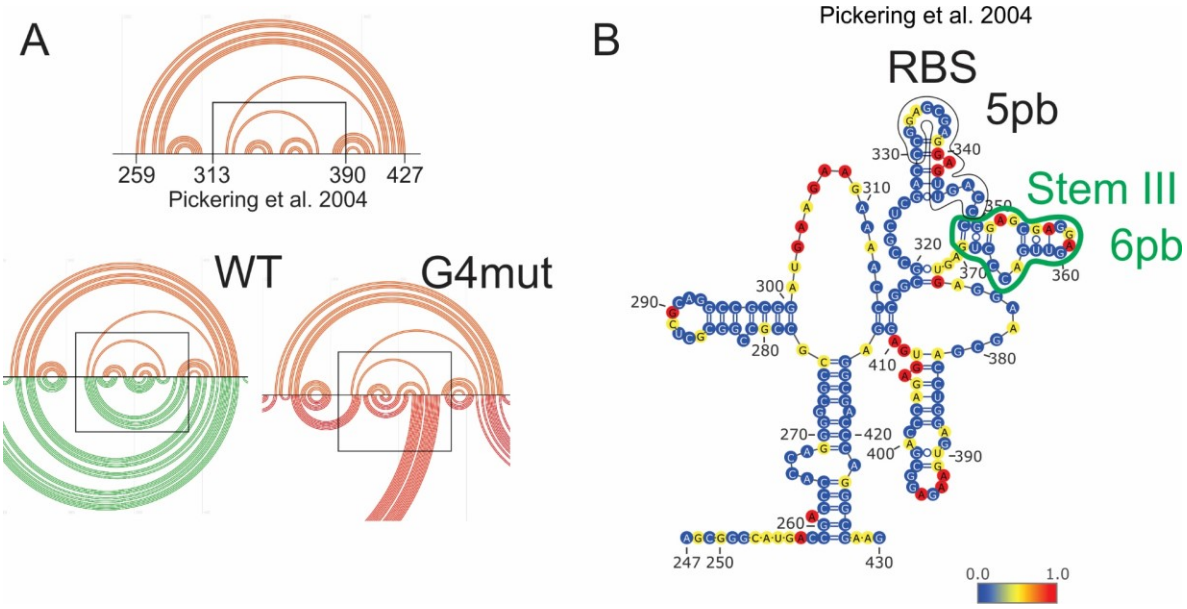

**Supplementary Figure S5**. Control of the bicistronic constructions' integrity.

Total RNA was extracted from HCT116 cells transfected either with the pRL-HL (lane 1), the pRL-BAG1wt-HL (lane 2) or the pRL-BAG1g4mut-HL (lane 3) construction and was migrated on a denaturing agarose gel along with a positive control of *in vitro* transcribed bicistronic RNA derived from pRL-BAG-1wt-HL (lane 4). The gel was transferred to a Hybond XL membrane and probed using the specific Fluc or Rluc probes listed in **Supplementary Table S3** (A) Northern blot using the Fluc specific probe. (B) The same membrane as in (A) was exposed for 64 h following stripping in order to confirm the removal of the Fluc probe. (C) Two exposure times of the same Northern blot membrane using the Rluc specific probe. Except for the smear in lane 4 that was caused by the presence of incomplete *in vitro* transcribed sequences, only 1 band is present for each construction at the same position for both probes, indicating that the bicistronic constructions are intact, that is no cryptic promoters or splicing sites are present and possess both luciferases on the same RNA strand.




**Supplementary Figure S6.** Secondary structure elucidated by SHAPE of the minimal IRES region of the BAG-1 5'UTR.

(A) WT (B) rG4mut (C) WT and IRESmutA, and (D) rG4mut and IRESmutA. The colors represent the normalised SHAPE reactivity of each nucleotide. Blue, non-reactive; yellow, reactive; and, red, highly reactive. The RBS, the Stem III region and the IRES mut A mutation are all highlighted.




**Supplementary Figure S7.** Comparison of the IRES secondary structure elucidated by SHAPE using the WT complete BAG-1 5'UTR sequence with the structure elucidated by Pickering *et al.* 2004.

(A) Arc-plots of the secondary structures: shown are the IRES secondary structure previously proposed by Pickering *et al.* 2004 (orange) and the WT (green) and rG4mut (red) structures, both of which were proposed in this work. The nucleotide positions are indicated, the boxed region corresponds to the RBS and Stem-loop III domains. Of the total of 42 bp of the structure proposed by Pickering et al., 7 are identical with the WT structure and 12 with the rG4mut structure elucidated in this work. (B) Proposed secondary structure of the IRES region of Pickering *et al.* with the RBS and the Stem III highlighted. The color of the nucleotide represent the normalised SHAPE reactivity, as measured in this study, Blue, non-reactive; yellow, reactive; and, red, highly reactive. The reactivities obtained in this work do not agree entirely with the structure proposed by Pickering *et al*. 2004.

**Supplementary Tables**

**Supplementary Table S1. Clinicopathological parameters of the CRC patients.**

| Parameters | Value |
|---|---|
| **Age, median years (range)** | 69 (50-85) |
| ***Sex*** | |
| **Male, no (%)** | 26 (55%) |
| **Female, no (%)** | 24 (46%) |
| ***Tumor localization*** | |
| **Left + rectum, no (%)** | 12 (26%) |
| **Right + transverse, no (%)** | 34 (72%) |
| **Polypes, no (%)** | 1 (2%) |
| ***Tumor stage (TNM)*** | |
| **Adenoma, no (%)** | 9 (19%) |
| **1, no (%)** | 8 (17%) |
| **2, no (%)** | 10 (21%) |
| **3, no (%)** | 10 (21%) |
| **4, no (%)** | 10 (21%) |

**Supplementary Table S2. Translation initiation efficiency of the start codons of the BAG-1 5'UTR.**

| Isoforms | Frame | Translation initiation context | Translation initiation efficiency[1] |
|---|---|---|---|
| BAG-1L | 3 | GGGC**CUG**G | 10.7 |
| BAG-1M | 3 | CCGCGG**AUG**AA | 60 |
| BAG-1S | 3 | GAAGAG**AUG**AA | 93 |
| AUG-254 | 1 | GCGGGC**AUG**AC | 103 |
| AUG-254mut | 1 | GCGGGC**ACG**AC | 1 |
| Reference Initiation context[2] | *NA* | GCCACC**AUG**GG | 100 |

1. Refs: (Diaz de Arce *et al.*, 2018 ; Noderer *et al.*, 2014)
2. Ref. (Kozak, 1987)

**Supplementary Table S3. List of primers and oligonucleotides used in this study.**

| Method | Name | Sequence 5'-3' |
|---|---|---|
| **Cloning** | 5'UTR BAG-1 WT **NheI** restriction site | **GCTAGC**AGGCCGGGGCGGGGCTGGGAAGTAGTCGG GCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCG CTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGC GGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCG GCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCC GGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGT GGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAG TACTGCCAGCGGGCATGACCGACCCACCAGGGGCG CCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAG AAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTT GACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAG CGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGA GGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGG AGGTGACCAGGGAGGAA**GCTAGC** |
| | 5'UTR BAG-1 G4mut **NheI** restriction site | **GCTAGC**AGGCCGAGACGAGACTGAGAAGTAGTCGA GCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCG CTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGC GGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCG GCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCC GGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGT GGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAG TACTGCCAGCGGGCATGACCGACCCACCAGGGGCG CCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAG AAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTT GACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAG CGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGA GGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGG AGGTGACCAGGGAGGAA**GCTAGC** |
| | 5'UTR BAG-1 WT **1Smut** **NheI** restriction site | **GCTAGC**AGGCCGGGGCGGGGCTGGGAAGTAGTCGG GCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCG CTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGC GGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCG GCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCC GGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGT GGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAG TACTGCCAGCGGGCATGACCGACCCACCAGGGGCG CCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAG AAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTT GACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAG CGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG GCGACCCAGGGCGAAGAG**AGG**AATCGGAGCCAGGA GGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGG AGGTGACCAGGGAGGAA**GCTAGC** |
| | 5'UTR BAG-1 G4mut **1Smut** **NheI** restriction site | **GCTAGC**AGGCCGAGACGAGACTGAGAAGTAGTCGA GCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCG CTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGC |

| | | |
|---|---|---|
| | | GGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCG<br>GCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCC<br>GGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGT<br>GGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAG<br>TACTGCCAGCGGGCATGACCGACCCACCAGGGGCG<br>CCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAG<br>AAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTT<br>GACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAG<br>CGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAG**AGG**AATCGGAGCCAGGA<br>GGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGG<br>AGGTGACCAGGGAGGAA**GCTAGC** |
| **Amplify 5'UTR** | BAG1 complete 5UTR fwd | TCAGTCAGA**GCTAGC**AGGCCG |
| | BAG1 complete 5UTR rev | AGTCAGT**GCTAGC**TTCCTCCCTGGTCACCTCC |
| **Mut start codon 1M** | Q5-BAG-1Mmut_Fwd | AGGCCGCGGAGGAAGAAGAAAACCCGGCG |
| | Q5-BAG-1Mmut_Rev | GCGAGCGCCGGCGGCGGC |
| **Mut start codon 1L** | Fwd_psiCHECK_Nhe1 | CGACTCACTATAGGCTAGCAGGCCG |
| | Rev_BAG1L_mut | CCGCGCTGAGCCCGGCCCGCACTTG |
| | Fwd_BAG1L_mut | CAAGTGCGGGCCGGGCTCAGCGCGG |
| | Rev_psiCHECK_Nhe1_end | GCCATGGTGGCTAGCGGTTCCTCCC |
| **Bicistronic construct** | Hpa1_Q5_F | GTTAACATGGAAGACGCCAAAAACATAAAGAAAGG<br>C |
| | Hpa1_Q5_R | GACGTCCTGTGGGCGGCG |
| | Not1_BAG1_WT_5UTR_f | AAGCACGCGGCCGCAGGCCGGGGCGGGGCTGGG |
| | Not1_BAG1_G4mut_5UTR_f | AAGCACGCGGCCGCAGGCCGAGACGAGACTGAG |
| | HpaI_BAG1_5UTR_r | CCGTTAACTTCCTCCCTGGTCACCTCC |
| **IRES mutations** | Q5_stm3mutA_F | GGAGTTGACCGGCAGTGAGGAAGCGACC |
| | Q5_stm3mutA_R | TCGCTCCGGGTCAACTCC |
| | Q5_IRESmutB_F | CGACCCGGAGGTTGGAGTTGACCCGGAG |
| | Q5_IRESmutB_R | AGCGGCGCCGGGTTTTCT |
| **qPCR** | BAG1_G_f | GGAGAGTAAAAGCCACAATAGAGCAG |
| | BAG1_G_r | CTGTCTTTGAAATTTTCTGGCAGGAT |
| | MRPL19_G_3_f | AAGGAGAAAAGTACTCCACATTCCAGAG |
| | MRPL19_G_3_r | TGGGTCAGCTGTAGTAACACGA |
| | SDHA_G_f | TGTTGATGGGAACAAGAGGGCA |
| | SDHA_G_r | GCCTACCACCACTGCATCAAAT |
| | YWHAZ_G_f | TCCCCAATGCTTCACAAGCAGA |
| | YWHAZ_G_r | TCTTGTCATCACCAGCGGCAA |
| | fLuc.q.F2 | GTGGGCAAGGTGGTGCCATT |
| | fLuc.q.R2 | AATCATAGGGCCGCGCACAC |
| | rLuc.q.F2 | AAGGGCCTCCACTTCAGCCA |
| | rLuc.q.R2 | TTCTTCAGCACGCGCTCCAC |
| | fLuc_pRL-HL.q.F1 | TGGCAGGTCTTCCCGACGAT |
| | fLuc_pRL-HL.q.R1 | ACACAACTCCTCCGCGCAAC |
| | rLuc_pRL-HL.q.F1 | ACATGGTAACGCGGCCTCTTC |
| | rLuc_pRL-HL.q.R1 | ACCAGATTTGCCTGATTTGCCCA |
| **ddPCR** | PUM1_global_for_1 | TGAGGTGTGCACCATGAAC |
| | PUM1_global_rev_1 | CAGAATGTGCTTGCCATAGG |
| | B2M.qref.F3 | ACTACACTGAATTCACCCCCACTGA |

| | | |
|---|---|---|
| | B2M.qref.R3 | GCTGCTTACATGTCTCGATCCCA |
| | MRPL19.qref.F1 | TCATCGTGGACAAGCACCGC |
| | MRPL19.qref.R1 | TCAGAGGATCTGTTCTTCCCCTTCG |
| | YWHAZ.qref.F2 | TGAAGAGTCATACAAAGACAGCACGC |
| | YWHAZ.qref.R2 | AGACAAAAGTTGGAAGGCCGGT |
| **Northern Blot probes** | Fluc_cds_pRL-HL | GGATCTCTCTGATTTTTCTTGCGTCGAG |
| | Rluc_cds_pRL-HL | CCATAAATAAGAAGAGGCCGCGTTACCA |
| **Creation of DNA templates for mRNA synthesis** | Q5_intercistron_WT_fwd | ATTGTAATACTCTAGAGGATCCCCCGGGCGAGCTC CCGCGGCCGCAGGCCGGGGCGG |
| | Q5_intercistron_G4mut_fwd | ATTGTAATACTCTAGAGGATCCCCCGGGCGAGCTC CCGCGGCCGCAGGCCGAGACGAG |
| | Q5_intercistron_rev | AGCTAAGAATTTCGTCATCGCTGAATACAGTTACA TTTCTAGAATTATTGTTCATTTTTGAG |
| | P1-Rluc | CGCCGTAATACGACTCACTATAGGGCTAGCCACCA TGACTTCGAAAG |
| | P1-s1-Rluc | CGCCGTAATACGACTCACTATAGGGAGTGGACTTC GGTCCACTCCCCTAGCCACCATGACTTCGAAAG |
| | P1-rev-Rluc | $(T)_{-60}$GGGAGCTCGCCCGGGGGATCC |
| | P2-BAG1 | CGCCGTAATACGACTCACTATAGGGAGGCCGGGGC GGGGCTGGGAAGTAG |
| | P2-BAG1G4mut | CGCCGTAATACGACTCACTATAGGGAGGCCGAGAC GAGACTGAGAAGTAG |
| | P2-S1-BAG1-short | CGCCGTAATACGACTCACTATAGGGAGTGGACTTC GGTCCACTCCCAGGCCGGGGCGGGGC |
| | P2-S1-BAG1G4mut | CGCCGTAATACGACTCACTATAGGGAGTGGACTTC GGTCCACTCCCAGGCCGAGACGAGACTGAGAAGTA G |
| | P3 | $(T)_{-60}$GAATAGAATGACACCTACTCAGAC |
| | P3-long | $(T)_{-60}$GAATAGAATGACACCTACTCAGACAATGCGA TGCAATTTC |
| | P4-Fluc | CGCCGTAATACGACTCACTATAGGGGAGGAAGTTA ACATGGAAGA |
| **SHAPE DNA template for *in vitro* trx** | T3_BAG1wt_fwd | AATTAACCCTCACTAAAGAGGCCGGGGCGGGGCTG GGA |
| | T3_BAG1G4mut_fwd | AATTAACCCTCACTAAAGAGGCCGAGACGAGACTG AGA |
| | Rluc_psicheck-2_rev | GCTCGGGGTCGTACACCTTG |
| | SHAPE_BAG1IRESmutA-R | GCTCGGGGTCGTACACCTTGGAAGCCATGGTGGCT AGCGGTTCCTCCCTGGTCACCTCCT |
| **SHAPE primers for RT** | Shape_BAG1_rev | 6-FAM GCTCGGGGTCGTACACCTTG |
| | Seq_BAG1_rev | NED GCTCGGGGTCGTACACCTTG |
| | Shape_BAG1no2_r | 6-FAM CGCCGGGTTTTCTTCTTCAT |
| | Seq_BAG1no2_r | NED CGCCGGGTTTTCTTCTTCAT |

**Supplementary Table S4. Sequences of all of the 5'UTRs tested, those of the rG4 and the start codons that were mutated, the SHAPE WT and the mutated sequences.**

| Construction DNA | G4 | Mutations | Sequence 5'-3' |
|---|---|---|---|
| BAG-1 complete 5'UTR | WT | WT | AGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG ACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG GAGGTGACCAGGGAGGAA |
| | | 1S mut | AGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG ACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG GCGACCCAGGGCGAAGAGAGGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG GAGGTGACCAGGGAGGAA |
| | | 1M mut | AGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC GCCGCCGGCGCTCGCAGGCCGCGGAGGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG ACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG GAGGTGACCAGGGAGGAA |

| | | | |
|---|---|---|---|
| | | 1L mut | AGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |
| | | Stem3mutA | AGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGCGAGGAGTTGACCGGCAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |
| | | Stem3mutB | AGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGGTTTGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |
| | G4mut | WT | AGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |

| | | | |
|---|---|---|---|
| | | 1Smut | AGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGAGGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |
| | | 1Mmut | AGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGAGGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |
| | | 1Lmut | AGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGAGGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |
| | | Stem3mutA | AGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGCGAGGAGTTGACCGGCAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |

| | | | |
|---|---|---|---|
| | | Stem3mutB | AGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGG<br>CGGTCAACAAGTGCGGGCCGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTG<br>GGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGT<br>CCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCC<br>GCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTG<br>ACCCGGAGGTTGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAG<br>GCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAG<br>GAGGTGACCAGGGAGGAA |
| **DNA template for *in vitro* transcription of RNA for SHAPE** | **G4** | **mutation IRES** | **Sequence 5'-3'** |
| BAG-1 5'UTR SHAPE T3 promoter, not transcribed excepted the last **G** Extra sequence for RT primer binding | WT | WT | AATTAACCCTCACTAAAGAGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGC<br>TCAGCTTCCATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGA<br>GGCGACCGGGAGCGGCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAG<br>CCCCCGGCCCAGCGTGGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGAC<br>CGACCCACCAGGGGCGCCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCG<br>ACCCGGAGCGAGGAGTTGACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAG<br>GCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAG<br>GAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAACCGCTAGCCACCATGGCTTCCAAGGTGTACGAC<br>CCCGAGC |
| | | Stem3mutA | AATTAACCCTCACTAAAGAGGCCGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGC<br>TCAGCTTCCATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGA<br>GGCGACCGGGAGCGGCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAG<br>CCCCCGGCCCAGCGTGGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGAC<br>CGACCCACCAGGGGCGCCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCG<br>ACCCGGAGCGAGGAGTTGACCCGGAGCGAGGAGTTGACCGGCAGTGAGGAAGCGACCTGGAGTGAAGAG<br>GCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAG<br>GAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAACCGCTAGCCACCATGGCTTCCAAGGTGTACGAC<br>CCCGAGC |

| | | | |
|---|---|---|---|
| | G4mut | WT | AATTAACCCTCACTAAAGAGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGC<br>TCAGCTTCCATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGA<br>GGCGACCGGGAGCGGCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAG<br>CCCCCGGCCCAGCGTGGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGAC<br>CGACCCACCAGGGGCGCCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCG<br>ACCCGGAGCGAGGAGTTGACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAG<br>GCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAG<br>GAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAACCGCTAGCCACCATGGCTTCCAAGGTGTACGAC<br>CCCGAGC |
| | | Stem3mutA | AATTAACCCTCACTAAAGAGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGC<br>TCAGCTTCCATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGA<br>GGCGACCGGGAGCGGCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAG<br>CCCCCGGCCCAGCGTGGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGAC<br>CGACCCACCAGGGGCGCCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCG<br>ACCCGGAGCGAGGAGTTGACCCGGAGCGAGGAGTTGACCGGCAGTGAGGAAGCGACCTGGAGTGAAGAG<br>GCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAG<br>GAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAACCGCTAGCCACCATGGCTTCCAAGGTGTACGAC<br>CCCGAGC |

**Supplementary Table S5. Sequences of the transfected mono- and bicistronic mRNAs.**

| DNA templates for mRNA transfection | Sequence 5'-3' |
|---|---|
| **Ctrl Monocistronic Rluc 1111bp**<br>T7 promoter, not transcribed except last 3 Gs<br>Rluc cds | CGCCGTAATACGACTCACTATAGGGCTAGCCACCATGACTTCGAAAGTTTATGATCCAGAACAAAGGAAACGGATG<br>ATAACTGGTCCGCAGTGGTGGGCCAGATGTAAACAAATGAATGTTCTTGATTCATTTATTAATTATTATGATTCAG<br>AAAAACATGCAGAAAATGCTGTTATTTTTTTACATGGTAACGCGGCCTCTTCTTATTTATGGCGACATGTTGTGCC<br>ACATATTGAGCCAGTAGCGCGGTGTATTATACCAGACCTTATTGGTATGGGCAAATCAGGCAAATCTGGTAATGGT<br>TCTTATAGGTTACTTGATCATTACAAATATCTTACTGCATGGTTTGAACTTCTTAATTTACCAAAGAAGATCATTT<br>TTGTCGGCCATGATTGGGGTGCTTGTTTGGCATTTCATTATAGCTATGAGCATCAAGATAAGATCAAAGCAATAGT<br>TCACGCTGAAAGTGTAGTAGATGTGATTGAATCATGGGATGAATGGCCTGATATTGAAGAAGATATTGCGTTGATC<br>AAATCTGAAGAAGGAGAAAAAATGGTTTTGGAGAATAACTTCTTCGTGGAAACCATGTTGCCATCAAAAATCATGA<br>GAAAGTTAGAACCAGAAGAATTTGCAGCATATCTTGAACCATTCAAAGAGAAAGGTGAAGTTCGTCGTCCAACATT<br>ATCATGGCCTCGTGAAATCCCGTTAGTAAAAGGTGGTAAACCTGACGTTGTACAAATTGTTAGGAATTATAATGCT<br>TATCTACGTGCAAGTGATGATTTACCAAAAATGTTTATTGAATCGGACCCAGGATTCTTTTCCAATGCTATTGTTG |

| | |
|---|---|
| | AAGGTGCCAAGAAGTTTCCTAATACTGAATTTGTCAAAGTAAAAGGTCTTCATTTTTCGCAAGAAGATGCACCTGA<br>TGAAATGGGAAAATATATCAAATCGTTCGTTGAGCGAGTTCTCAAAAATGAACAATAATTCTAGAAATGTAACTGT<br>ATTCAGCGATGACGAAATTCTTAGCTATTGTAATACTCTAGAGGATCCCCCGGGCGAGCTCCCTTTTTTTTTTTTT<br>TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT |
| **Ctrl Monocistronic Fluc 1950bp**<br>Fluc cds | CGCCGTAATACGACTCACTATAGGGGAGGAAGTTAACATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCA<br>TTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAA<br>CAATTGCTTTTACAGATGCACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCCGTTCGGTTGGC<br>AGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACTCTCTTCAATTCTTT<br>ATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAGTTGCGCCCGCGAACGACATTTATAATGAACGTGAATTGC<br>TCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTTTGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCA<br>AAAAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAAACGGATTACCAGGGATTTCAGTCGATGTAC<br>ACGTTCGTCACATCTCATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGATCGTGACAAAA<br>CAATTGCACTGATAATGAATTCCTCTGGATCTACTGGGTTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTG<br>CGTCAGATTCTCGCATGCCAGAGATCCTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTAAGTGTTGTT<br>CCATTCCATCACGGTTTTGGAATGTTTACTACACTCGGATATTTGATATGTGGATTTCGAGTCGTCTTAATGTATA<br>GATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTATT<br>TTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCA<br>CCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCA<br>CTGAGACTACATCAGCTATTCTGATTACACCCGAGGGGGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATT<br>TTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGA<br>GGACCTATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGATGGATGGCTAC<br>ATTCTGGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCATAGTTGACCGCTTGAAGTCTTTAATTAAATA<br>CAAAGGATATCAGGTGGCCCCCGCTGAATTGGAATCGATATTGTTACAACACCCCAACATCTTCGACGCGGGCGTG<br>GCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGG<br>AAAAAGAGATCGTGGATTACGTGGCCAGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGA<br>CGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCCAAGAAGGGC<br>GGAAAGTCCAAATTGTAAGGATCCGGGCCCTATTCTATAGTGTCACCTAAATGCTAGAGCTCGCTGATCAGCCTCG<br>ACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTC<br>CCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTTTTTTTTTT<br>TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT |
| **Monocistronic Fluc WT 2445bp** | CGCCGTAATACGACTCACTATAGGGAGGCC**GGGG**C**GGGG**CT**GGG**AAGTAGTC**GGG**C**GGGG**TTGTGAGACGCCGCGC<br>TCAGCTTCCATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACC<br>GGGAGCGGCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCG<br>TGGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCCGCC<br>GCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTGACCCGGAGCG<br>AGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGA<br>GATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAAGTTAAC<br>ATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAAC |

| | |
|---|---|
| | TGCATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTACAGATGCACATATCGAGGTGAACAT<br>CACGTACGCGGAATACTTCGAAATGTCCGTTCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCAC<br>AGAATCGTCGTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAG<br>TTGCGCCCGCGAACGACATTTATAATGAACGTGAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTT<br>TGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATAATCCAGAAAATTATTATCATG<br>GATTCTAAAACGGATTACCAGGGATTTCAGTCGATGTACACGTTCGTCACATCTCATCTACCTCCCGGTTTTAATG<br>AATACGATTTTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAATGAATTCCTCTGGATCTACTGG<br>GTTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCATGCCAGAGATCCTATTTTTGGC<br>AATCAAATCATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACTACACTCG<br>GATATTTGATATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGA<br>TTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATAC<br>GATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAAC<br>GCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGG<br>GGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAA<br>ACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGAGGACCTATGATTATGTCCGGTTATGTAAACAATCCGG<br>AAGCGACCAACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACA<br>CTTCTTCATAGTTGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTGGCCCCCGCTGAATTGGAATCG<br>ATATTGTTACAACACCCCAACATCTTCGACGCGGGCGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCG<br>CCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGATTACGTGGCCAGTCAAGTAAC<br>AACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCA<br>AGAAAAATCAGAGAGATCCTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGTAAGGATCCGGGCCCTATTCTA<br>TAGTGTCACCTAAATGCTAGAGCTCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCC<br>CCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATC<br>GCATTGTCTGAGTAGGTGTCATTCTATTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT<br>TTTTTTTTTTTT |
| **Monocistronic Fluc G4mut 2445bp** | CGCCGTAATACGACTCACTATAGGGAGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGC<br>TCAGCTTCCATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACC<br>GGGAGCGGCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCG<br>TGGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCCGCC<br>GCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTGACCCGGAGCG<br>AGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGA<br>GATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAAGTTAAC<br>ATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAAC<br>TGCATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTACAGATGCACATATCGAGGTGAACAT<br>CACGTACGCGGAATACTTCGAAATGTCCGTTCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCAC<br>AGAATCGTCGTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAG<br>TTGCGCCCGCGAACGACATTTATAATGAACGTGAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTT<br>TGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATAATCCAGAAAATTATTATCATG |

| | |
|---|---|
| | GATTCTAAAACGGATTACCAGGGATTTCAGTCGATGTACACGTTCGTCACATCTCATCTACCTCCCGGTTTTAATG<br>AATACGATTTTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAATGAATTCCTCTGGATCTACTGG<br>GTTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCATGCCAGAGATCCTATTTTTGGC<br>AATCAAATCATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACTACACTCG<br>GATATTTGATATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGA<br>TTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATAC<br>GATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAAC<br>GCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGG<br>GGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAA<br>ACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGAGGACCTATGATTATGTCCGGTTATGTAAACAATCCGG<br>AAGCGACCAACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACA<br>CTTCTTCATAGTTGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTGGCCCCCGCTGAATTGGAATCG<br>ATATTGTTACAACACCCCAACATCTTCGACGCGGGCGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCG<br>CCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGATTACGTGGCCAGTCAAGTAAC<br>AACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCA<br>AGAAAAATCAGAGAGATCCTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGTAAGGATCCGGGCCCTATTCTA<br>TAGTGTCACCTAAATGCTAGAGCTCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCC<br>CCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATC<br>GCATTGTCTGAGTAGGTGTCATTCTATTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT<br>TTTTTTTTTTTTT |
| **Monocistronic Hairpin Fluc WT 2466bp**<br>Hairpin sequence | CGCCGTAATACGACTCACTATAGGGAGTGGACTTCGGTCCACTCCCAGGCCGGGGCGGGGCTGGGAAGTAGTCGGG<br>CGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGG<br>CGCGGAGACCGCGAGGCGACCGGGAGCGGCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCA<br>GTCGGAGCCCCCGGCCCAGCGTGGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGAC<br>CGACCCACCAGGGGCGCCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGA<br>GCGAGGAGTTGACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGA<br>GGAGGCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGGAG<br>GTGACCAGGGAGGAAGTTAACATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGG<br>ATGGAACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTACAGA<br>TGCACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCCGTTCGGTTGGCAGAAGCTATGAAACGA<br>TATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGGCG<br>CGTTATTTATCGGAGTTGCAGTTGCGCCCGCGAACGACATTTATAATGAACGTGAATTGCTCAACAGTATGAACAT<br>TTCGCAGCCTACCGTAGTGTTTGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATA<br>ATCCAGAAAATTATTATCATGGATTCTAAAACGGATTACCAGGGATTTCAGTCGATGTACACGTTCGTCACATCTC<br>ATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAAT<br>GAATTCCTCTGGATCTACTGGGTTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCAT<br>GCCAGAGATCCTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTT<br>TTGGAATGTTTACTACACTCGGATATTTGATATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCT |

| | |
|---|---|
| | GTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAAA<br>AGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAG<br>TCGGGGAAGCGGTTGCAAAACGCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGC<br>TATTCTGATTACACCCGAGGGGGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTT<br>GTGGATCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGAGGACCTATGATTATGT<br>CCGGTTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGC<br>TTACTGGGACGAAGACGAACACTTCTTCATAGTTGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTG<br>GCCCCCGCTGAATTGGAATCGATATTGTTACAACACCCCAACATCTTCGACGCGGGCGTGGCAGGTCTTCCCGACG<br>ATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGA<br>TTACGTGGCCAGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAAGGT<br>CTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGT<br>AAGGATCCGGGCCCTATTCTATAGTGTCACCTAAATGCTAGAGCTCGCTGATCAGCCTCGACTGTGCCTTCTAGTT<br>GCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTA<br>ATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTTTTTTTTTTTTTTTTTTTTTTTT<br>TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT |
| **Monocistronic Hairpin Fluc G4mut 2466bp** | CGCCGTAATACGACTCACTATAGGGAGTGGACTTCGGTCCACTCCCAGGCCGAGACGAGACTGAGAAGTAGTCGAG<br>CGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGG<br>CGCGGAGACCGCGAGGCGACCGGGAGCGGCTGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCA<br>GTCGGAGCCCCCGGCCCAGCGTGGTCCGCCTCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGAC<br>CGACCCACCAGGGGCGCCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGA<br>GCGAGGAGTTGACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGA<br>GGAGGCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGGAG<br>GTGACCAGGGAGGAAGTTAACATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGG<br>ATGGAACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTACAGA<br>TGCACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCCGTTCGGTTGGCAGAAGCTATGAAACGA<br>TATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGGCG<br>CGTTATTTATCGGAGTTGCAGTTGCGCCCGCGAACGACATTTATAATGAACGTGAATTGCTCAACAGTATGAACAT<br>TTCGCAGCCTACCGTAGTGTTTGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATA<br>ATCCAGAAAATTATTATCATGGATTCTAAAACGGATTACCAGGGATTTCAGTCGATGTACACGTTCGTCACATCTC<br>ATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAAT<br>GAATTCCTCTGGATCTACTGGGTTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCAT<br>GCCAGAGATCCTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTT<br>TTGGAATGTTTACTACACTCGGATATTTGATATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCT<br>GTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAAA<br>AGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAG<br>TCGGGGAAGCGGTTGCAAAACGCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGC<br>TATTCTGATTACACCCGAGGGGGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTT<br>GTGGATCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGAGGACCTATGATTATGT |

| | |
|---|---|
| | CCGGTTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGC<br>TTACTGGGACGAAGACGAACACTTCTTCATAGTTGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTG<br>GCCCCCGCTGAATTGGAATCGATATTGTTACAACACCCCAACATCTTCGACGCGGGCGTGGCAGGTCTTCCCGACG<br>ATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGA<br>TTACGTGGCCAGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAAGGT<br>CTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGT<br>AAGGATCCGGGCCCTATTCTATAGTGTCACCTAAATGCTAGAGCTCGCTGATCAGCCTCGACTGTGCCTTCTAGTT<br>GCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTA<br>ATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTTTTTTTTTTTTTTTTTTTTTTTT<br>TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT |
| **Bicistronic WT 3479bp** | CGCCGTAATACGACTCACTATAGGGCTAGCCACCATGACTTCGAAAGTTTATGATCCAGAACAAAGGAAACGGATG<br>ATAACTGGTCCGCAGTGGTGGGCCAGATGTAAACAAATGAATGTTCTTGATTCATTTATTAATTATTATGATTCAG<br>AAAAACATGCAGAAAATGCTGTTATTTTTTTACATGGTAACGCGGCCTCTTCTTATTTATGGCGACATGTTGTGCC<br>ACATATTGAGCCAGTAGCGCGGTGTATTATACCAGACCTTATTGGTATGGGCAAATCAGGCAAATCTGGTAATGGT<br>TCTTATAGGTTACTTGATCATTACAAATATCTTACTGCATGGTTTGAACTTCTTAATTTACCAAAGAAGATCATTT<br>TTGTCGGCCATGATTGGGGTGCTTGTTTGGCATTTCATTATAGCTATGAGCATCAAGATAAGATCAAAGCAATAGT<br>TCACGCTGAAAGTGTAGTAGATGTGATTGAATCATGGGATGAATGGCCTGATATTGAAGAAGATATTGCGTTGATC<br>AAATCTGAAGAAGGAGAAAAAATGGTTTTGGAGAATAACTTCTTCGTGGAAACCATGTTGCCATCAAAAATCATGA<br>GAAAGTTAGAACCAGAAGAATTTGCAGCATATCTTGAACCATTCAAAGAGAAAGGTGAAGTTCGTCGTCCAACATT<br>ATCATGGCCTCGTGAAATCCCGTTAGTAAAAGGTGGTAAACCTGACGTTGTACAAATTGTTAGGAATTATAATGCT<br>TATCTACGTGCAAGTGATGATTTACCAAAAATGTTTATTGAATCGGACCCAGGATTCTTTTCCAATGCTATTGTTG<br>AAGGTGCCAAGAAGTTTCCTAATACTGAATTTGTCAAAGTAAAAGGTCTTCATTTTTCGCAAGAAGATGCACCTGA<br>TGAAATGGGAAAATATATCAAATCGTTCGTTGAGCGAGTTCTCAAAAATGAACAATAATTCTAGAAATGTAACTGT<br>ATTCAGCGATGACGAAATTCTTAGCTATTGTAATACTCTAGAGGATCCCCCGGGCGAGCTCCCGCGGCCGCAGGCC<br>GGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGGCGGTCAACAAGT<br>GCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTGGGTTCCCGGCTGCGCGCCC<br>TTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGTCCGCCTCCCTCTCGGCGTCCACCTGC<br>CCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAG<br>AAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTGACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCT<br>GGAGTGAAGAGGCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGA<br>CGAGGAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAAGTTAACATGGAAGACGCCAAAAACATAAAGAAAGGC<br>CCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGG<br>TTCCTGGAACAATTGCTTTTACAGATGCACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCCGT<br>TCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACTCTCTT<br>CAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAGTTGCGCCCGCGAACGACATTTATAATGAAC<br>GTGAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTTTGTTTCCAAAAAGGGGTTGCAAAAAATTTT<br>GAACGTGCAAAAAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAAACGGATTACCAGGGATTTCAG<br>TCGATGTACACGTTCGTCACATCTCATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGATC |

| | |
|---|---|
| | GTGACAAAACAATTGCACTGATAATGAATTCCTCTGGATCTACTGGGTTACCTAAGGGTGTGGCCCTTCCGCATAG<br>AACTGCCTGCGTCAGATTCTCGCATGCCAGAGATCCTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTA<br>AGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACTACACTCGGATATTTGATATGTGGATTTCGAGTCGTCT<br>TAATGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACC<br>AACCCTATTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATTGCTTCT<br>GGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGCTTCCATCTTCCAGGGATACGACAAGGAT<br>ATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGGGGATGATAAACCGGGCGCGGTCGGTAAAGT<br>TGTTCCATTTTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTA<br>TGTGTCAGAGGACCTATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGATG<br>GATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCATAGTTGACCGCTTGAAGTCTTT<br>AATTAAATACAAAGGATATCAGGTGGCCCCCGCTGAATTGGAATCGATATTGTTACAACACCCCAACATCTTCGAC<br>GCGGGCGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGA<br>CGATGACGGAAAAAGAGATCGTGGATTACGTGGCCAGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGT<br>GTTTGTGGACGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCC<br>AAGAAGGGCGGAAAGTCCAAATTGTAAGGATCCGGGCCCTATTCTATAGTGTCACCTAAATGCTAGAGCTCGCTGA<br>TCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAG<br>GTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCT<br>TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT |
| **Bicistronic G4mut 3479bp** | CGCCGTAATACGACTCACTATAGGGCTAGCCACCATGACTTCGAAAGTTTATGATCCAGAACAAAGGAAACGGATG<br>ATAACTGGTCCGCAGTGGTGGGCCAGATGTAAACAAATGAATGTTCTTGATTCATTTATTAATTATTATGATTCAG<br>AAAAACATGCAGAAAATGCTGTTATTTTTTTACATGGTAACGCGGCCTCTTCTTATTTATGGCGACATGTTGTGCC<br>ACATATTGAGCCAGTAGCGCGGTGTATTATACCAGACCTTATTGGTATGGGCAAATCAGGCAAATCTGGTAATGGT<br>TCTTATAGGTTACTTGATCATTACAAATATCTTACTGCATGGTTTGAACTTCTTAATTTACCAAAGAAGATCATTT<br>TTGTCGGCCATGATTGGGGTGCTTGTTTGGCATTTCATTATAGCTATGAGCATCAAGATAAGATCAAAGCAATAGT<br>TCACGCTGAAAGTGTAGTAGATGTGATTGAATCATGGGATGAATGGCCTGATATTGAAGAAGATATTGCGTTGATC<br>AAATCTGAAGAAGGAGAAAAAATGGTTTTGGAGAATAACTTCTTCGTGGAAACCATGTTGCCATCAAAAATCATGA<br>GAAAGTTAGAACCAGAAGAATTTGCAGCATATCTTGAACCATTCAAAGAGAAAGGTGAAGTTCGTCGTCCAACATT<br>ATCATGGCCTCGTGAAATCCCGTTAGTAAAAGGTGGTAAACCTGACGTTGTACAAATTGTTAGGAATTATAATGCT<br>TATCTACGTGCAAGTGATGATTTACCAAAAATGTTTATTGAATCGGACCCAGGATTCTTTTCCAATGCTATTGTTG<br>AAGGTGCCAAGAAGTTTCCTAATACTGAATTTGTCAAAGTAAAAGGTCTTCATTTTTCGCAAGAAGATGCACCTGA<br>TGAAATGGGAAAATATATCAAATCGTTCGTTGAGCGAGTTCTCAAAAATGAACAATAATTCTAGAAATGTAACTGT<br>ATTCAGCGATGACGAAATTCTTAGCTATTGTAATACTCTAGAGGATCCCCCGGGCGAGCTCCCGCGGCCGCAGGCC<br>GAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTCAGCTTCCATCGCTGGGCGGTCAACAAGT<br>GCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGCTGGGTTCCCGGCTGCGCGCCC<br>TTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGTCCGCCTCCCTCTCGGCGTCCACCTGC<br>CCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCCGCCGCCGGCGCTCGCAGGCCGCGGATGAAGAAG<br>AAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTGACCCGGAGCGAGGAGTTGACCCTGAGTGAGGAAGCGACCT<br>GGAGTGAAGAGGCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGAGATGAATCGGAGCCAGGAGGTGACCCGGGA |

| | |
|---|---|
| | CGAGGAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAAGTTAACATGGAAGACGCCAAAAACATAAAGAAAGGC<br>CCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGG<br>TTCCTGGAACAATTGCTTTTACAGATGCACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCCGT<br>TCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACTCTCTT<br>CAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAGTTGCGCCCGCGAACGACATTTATAATGAAC<br>GTGAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTTTGTTTCCAAAAAGGGGTTGCAAAAAATTTT<br>GAACGTGCAAAAAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAAACGGATTACCAGGGATTTCAG<br>TCGATGTACACGTTCGTCACATCTCATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGATC<br>GTGACAAAACAATTGCACTGATAATGAATTCCTCTGGATCTACTGGGTTACCTAAGGGTGTGGCCCTTCCGCATAG<br>AACTGCCTGCGTCAGATTCTCGCATGCCAGAGATCCTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTA<br>AGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACTACACTCGGATATTTGATATGTGGATTTCGAGTCGTCT<br>TAATGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACC<br>AACCCTATTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATTGCTTCT<br>GGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGCTTCCATCTTCCAGGGATACGACAAGGAT<br>ATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGGGGATGATAAACCGGGCGCGGTCGGTAAAGT<br>TGTTCCATTTTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTA<br>TGTGTCAGAGGACCTATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGATG<br>GATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCATAGTTGACCGCTTGAAGTCTTT<br>AATTAAATACAAAGGATATCAGGTGGCCCCCGCTGAATTGGAATCGATATTGTTACAACACCCCAACATCTTCGAC<br>GCGGGCGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGA<br>CGATGACGGAAAAAGAGATCGTGGATTACGTGGCCAGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGT<br>GTTTGTGGACGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCC<br>AAGAAGGGCGGAAAGTCCAAATTGTAAGGATCCGGGCCCTATTCTATAGTGTCACCTAAATGCTAGAGCTCGCTGA<br>TCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAG<br>GTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCT<br>TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT |
| **Bicistronic Hairpin WT 3500pb** | CGCCGTAATACGACTCACTATAGGGAGTGGACTTCGGTCCACTCCCCTAGCCACCATGACTTCGAAAGTTTATGAT<br>CCAGAACAAAGGAAACGGATGATAACTGGTCCGCAGTGGTGGGCCAGATGTAAACAAATGAATGTTCTTGATTCAT<br>TTATTAATTATTATGATTCAGAAAAACATGCAGAAAATGCTGTTATTTTTTTACATGGTAACGCGGCCTCTTCTTA<br>TTTATGGCGACATGTTGTGCCACATATTGAGCCAGTAGCGCGGTGTATTATACCAGACCTTATTGGTATGGGCAAA<br>TCAGGCAAATCTGGTAATGGTTCTTATAGGTTACTTGATCATTACAAATATCTTACTGCATGGTTTGAACTTCTTA<br>ATTTACCAAAGAAGATCATTTTTGTCGGCCATGATTGGGGTGCTTGTTTGGCATTTCATTATAGCTATGAGCATCA<br>AGATAAGATCAAAGCAATAGTTCACGCTGAAAGTGTAGTAGATGTGATTGAATCATGGGATGAATGGCCTGATATT<br>GAAGAAGATATTGCGTTGATCAAATCTGAAGAAGGAGAAAAAATGGTTTTGGAGAATAACTTCTTCGTGGAAACCA<br>TGTTGCCATCAAAAATCATGAGAAAGTTAGAACCAGAAGAATTTGCAGCATATCTTGAACCATTCAAAGAGAAAGG<br>TGAAGTTCGTCGTCCAACATTATCATGGCCTCGTGAAATCCCGTTAGTAAAAGGTGGTAAACCTGACGTTGTACAA<br>ATTGTTAGGAATTATAATGCTTATCTACGTGCAAGTGATGATTTACCAAAAATGTTTATTGAATCGGACCCAGGAT<br>TCTTTTCCAATGCTATTGTTGAAGGTGCCAAGAAGTTTCCTAATACTGAATTTGTCAAAGTAAAAGGTCTTCATTT |

| | |
|---|---|
| | TTCGCAAGAAGATGCACCTGATGAAATGGGAAAATATATCAAATCGTTCGTTGAGCGAGTTCTCAAAAATGAACAA<br>TAATTCTAGAAATGTAACTGTATTCAGCGATGACGAAATTCTTAGCTATTGTAATACTCTAGAGGATCCCCCGGGC<br>GAGCTCCCGCGGCCGCAGGCCGGGGGCGGGGCTGGGAAGTAGTCGGGCGGGGTTGTGAGACGCCGCGCTCAGCTTCC<br>ATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGC<br>TGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGTCCGCC<br>TCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCCGCCGCCGGCGCT<br>CGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTGACCCGGAGCGAGGAGTTGA<br>CCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGAGATGAATCG<br>GAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAAGTTAACATGGAAGAC<br>GCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGCATAAGG<br>CTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTACAGATGCACATATCGAGGTGAACATCACGTACGC<br>GGAATACTTCGAAATGTCCGTTCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTC<br>GTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAGTTGCGCCCG<br>CGAACGACATTTATAATGAACGTGAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTTTGTTTCCAA<br>AAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAA<br>ACGGATTACCAGGGATTTCAGTCGATGTACACGTTCGTCACATCTCATCTACCTCCCGGTTTTAATGAATACGATT<br>TTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAATGAATTCCTCTGGATCTACTGGGTTACCTAA<br>GGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCATGCCAGAGATCCTATTTTTGGCAATCAAATC<br>ATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACTACACTCGGATATTTGA<br>TATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTACAAAAT<br>TCAAAGTGCGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACGATTTATCT<br>AATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGCTTCCATC<br>TTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGGGGATGATAA<br>ACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAAACGCTGGGC<br>GTTAATCAGAGAGGCGAATTATGTGTCAGAGGACCTATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACCA<br>ACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCAT<br>AGTTGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTGGCCCCCGCTGAATTGGAATCGATATTGTTA<br>CAACACCCCAACATCTTCGACGCGGGCGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCGCCGCCGTTG<br>TTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGATTACGTGGCCAGTCAAGTAACAACCGCGAA<br>AAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATC<br>AGAGAGATCCTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGTAAGGATCCGGGCCCTATTCTATAGTGTCAC<br>CTAAATGCTAGAGCTCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCG<br>TGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCT<br>GAGTAGGTGTCATTCTATTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT<br>TTTT |

| | |
|---|---|
| **BicistronicHairpin G4mut 3500pb** | CGCCGTAATACGACTCACTATAGGGAGTGGACTTCGGTCCACTCCCCTAGCCACCATGACTTCGAAAGTTTATGAT<br>CCAGAACAAAGGAAACGGATGATAACTGGTCCGCAGTGGTGGGCCAGATGTAAACAAATGAATGTTCTTGATTCAT<br>TTATTAATTATTATGATTCAGAAAAACATGCAGAAAATGCTGTTATTTTTTTACATGGTAACGCGGCCTCTTCTTA<br>TTTATGGCGACATGTTGTGCCACATATTGAGCCAGTAGCGCGGTGTATTATACCAGACCTTATTGGTATGGGCAAA<br>TCAGGCAAATCTGGTAATGGTTCTTATAGGTTACTTGATCATTACAAATATCTTACTGCATGGTTTGAACTTCTTA<br>ATTTACCAAAGAAGATCATTTTTGTCGGCCATGATTGGGGTGCTTGTTTGGCATTTCATTATAGCTATGAGCATCA<br>AGATAAGATCAAAGCAATAGTTCACGCTGAAAGTGTAGTAGATGTGATTGAATCATGGGATGAATGGCCTGATATT<br>GAAGAAGATATTGCGTTGATCAAATCTGAAGAAGGAGAAAAAATGGTTTTGGAGAATAACTTCTTCGTGGAAACCA<br>TGTTGCCATCAAAAATCATGAGAAAGTTAGAACCAGAAGAATTTGCAGCATATCTTGAACCATTCAAAGAGAAAGG<br>TGAAGTTCGTCGTCCAACATTATCATGGCCTCGTGAAATCCCGTTAGTAAAAGGTGGTAAACCTGACGTTGTACAA<br>ATTGTTAGGAATTATAATGCTTATCTACGTGCAAGTGATGATTTACCAAAAATGTTTATTGAATCGGACCCAGGAT<br>TCTTTTCCAATGCTATTGTTGAAGGTGCCAAGAAGTTTCCTAATACTGAATTTGTCAAAGTAAAAGGTCTTCATTT<br>TTCGCAAGAAGATGCACCTGATGAAATGGGAAAATATATCAAATCGTTCGTTGAGCGAGTTCTCAAAAATGAACAA<br>TAATTCTAGAAATGTAACTGTATTCAGCGATGACGAAATTCTTAGCTATTGTAATACTCTAGAGGATCCCCCGGGC<br>GAGCTCCCGCGGCCGCAGGCCGAGACGAGACTGAGAAGTAGTCGAGCGAGGTTGTGAGACGCCGCGCTCAGCTTCC<br>ATCGCTGGGCGGTCAACAAGTGCGGGCCTGGCTCAGCGCGGGGGGGCGCGGAGACCGCGAGGCGACCGGGAGCGGC<br>TGGGTTCCCGGCTGCGCGCCCTTCGGCCAGGCCGGGAGCCGCGCCAGTCGGAGCCCCCGGCCCAGCGTGGTCCGCC<br>TCCCTCTCGGCGTCCACCTGCCCGGAGTACTGCCAGCGGGCATGACCGACCCACCAGGGGCGCCGCCGCCGGCGCT<br>CGCAGGCCGCGGATGAAGAAGAAAACCCGGCGCCGCTCGACCCGGAGCGAGGAGTTGACCCGGAGCGAGGAGTTGA<br>CCCTGAGTGAGGAAGCGACCTGGAGTGAAGAGGCGACCCAGAGTGAGGAGGCGACCCAGGGCGAAGAGATGAATCG<br>GAGCCAGGAGGTGACCCGGGACGAGGAGTCGACCCGGAGCGAGGAGGTGACCAGGGAGGAAGTTAACATGGAAGAC<br>GCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGCATAAGG<br>CTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTACAGATGCACATATCGAGGTGAACATCACGTACGC<br>GGAATACTTCGAAATGTCCGTTCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTC<br>GTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAGTTGCGCCCG<br>CGAACGACATTTATAATGAACGTGAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTTTGTTTCCAA<br>AAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAA<br>ACGGATTACCAGGGATTTCAGTCGATGTACACGTTCGTCACATCTCATCTACCTCCCGGTTTTAATGAATACGATT<br>TTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAATGAATTCCTCTGGATCTACTGGGTTACCTAA<br>GGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCATGCCAGAGATCCTATTTTTGGCAATCAAATC<br>ATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACTACACTCGGATATTTGA<br>TATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTACAAAAT<br>TCAAAGTGCGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACGATTTATCT<br>AATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGCTTCCATC<br>TTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGGGGATGATAA<br>ACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAAACGCTGGGC<br>GTTAATCAGAGAGGCGAATTATGTGTCAGAGGACCTATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACCA<br>ACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCAT |

| | |
|---|---|
| | AGTTGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTGGCCCCCGCTGAATTGGAATCGATATTGTTA<br>CAACACCCCAACATCTTCGACGCGGGCGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCGCCGCCGTTG<br>TTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGATTACGTGGCCAGTCAAGTAACAACCGCGAA<br>AAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATC<br>AGAGAGATCCTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGTAAGGATCCGGGCCCTATTCTATAGTGTCAC<br>CTAAATGCTAGAGCTCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCG<br>TGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCT<br>GAGTAGGTGTCATTCTATTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT<br>TTTT |

**ANNEXE 6 Tableau A2 Banque de données sur les G4**

| Base de données | Acide nucléique | Source des données | Génome | Accessibilité web | Référence |
|---|---|---|---|---|---|
| QuadDB | ADN | Prédiction motif canonique | *Homo sapiens* | URL n'existe plus *http://www.quadruplex.org* | Wong, 2010 |
| Greglist | ADN | Prédiction motif canonique | *Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus* | URL n'existe plus *http://tubic.tju.edu.cn/greglist* | Zhang, 2008 |
| Quadbase | ADN | Prédiction motif canonique | Promoteurs de *Pan troglodytes, Mus musculus, Rattus norvegicus* et 146 microbes | *http://quadbase.igib.res.in/* | Yadav, 2008 |
| Quadbase2 | ADN | Prédiction motif canonique | 178 génomes eucaryotes dans le module EuQuad et 1719 génomes prokaryotes dans le module ProQuad | *http://quadbase.igib.res.in/* | Dhapola, 2016 |
| GRSDB2 et GRS_UTRdb | pré-ARNm | Prédiction motifs (QGRS) | *Homo sapiens, Mus musculus, Drosophila melanogaster, Rattus norvegicus, Caernohabditis elegans, Gallus gallus, Bos taurus, Danio Rerio* | *http://bioinformatics.ramapo.edu/GRSDB2/* et *http://bioinformatics.ramapo.edu/GQRS/* URL vers GRS-UTRdb non fonctionnel | Kikin, 2008 |
| Non-B DB | ADN | Prédiction motif canonique | *Homo sapiens, Pan troglodytes, Canis lupus, Macaca mulatta Mus musculus, Pongo abelii, Rattus norvegicus, Bos taurus, Sus scrofa, Equus caballus, Ornithorhynchus anatinus* et *Arabidopsis thaliana* | *http://nonb.abcc.ncifcrf.gov* | Cer, 2013 |

| Base de données | Acide nucléique | Source des données | Génome | Accessibilité web | Référence |
|---|---|---|---|---|---|
| G4RNA | ARN | Expérimentations, littérature scientifique | 334 séquences provenant du transcriptome *Homo sapiens* et de séquences artificielles rapportées | *http://scottgroup.med.usherbrooke.ca/G4RNA/* | Garant, 2015 |
| G4-seq | ADN | Expérimentation-séquençage | Culture primaire de Lymphocytes B humains | Non, résultats disponibles en Matériel supplémentaire | Chambers, 2015 |
| rG4-seq | ARN | Expérimentation-séquençage | ARN polyadénylés purifiés de cellules humaines HeLa | Non, résultats disponibles en Matériel supplémentaire | Kwok, 2016 |
| G4Hunter (392lit) | ADN | Expérimentation *in vitro* et prédictions (score G4H) | 392 séquences évaluées expérimentalement et publiées, génome mitochondrial humain (16.6 kb) et génomes complets : *Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Plasmodium falciparum, Escherichia coli, Arabidopsis thaliana et Dictyostelium discoideum* | Non, résultats disponibles en Matériel supplémentaire | Bedrat, 2016 |
| G4 génomes viraux | ADN/ARN selon le virus | Prédiction motif canonique | 7 familles virales et leurs différentes souches | *http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus* | Lavezzo, 2018 |
| G4-Seq Multiple species | ADN | Expérimentation-séquençage (méthode G4-Seq améliorée) | ADN génomique purifié, 12 espèces : *Homo sapiens (HEK-293T), Mus musculus, Danio Rerio, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Leishmania major, Trypanosoma brucei, Plasmodium falciparum, Arabidopsis thaliana, Rhodobacter sphaeroides, Escherichia coli* | Non, résultats disponibles en Matériel supplémentaire | Marsico, 2019 |

## ANNEXE 7 Figure 44 et Figure 45

### Légende Figure 44 et Figure 45

Les nucléotides en bleu sont ceux qui ont une absence ou une faible réactivité SHAPE. Les nucléotides en jaune sont ceux présentant une réactivité intermédiaire et les nucléotides en rouge une réactivité élevée. Les nucléotides en gris sont ceux dont la réactivité ne peut pas être déterminée puisqu'ils forment le site d'hybridation de l'amorce pour l'étape d'extension d'amorce. Les positions des nucléotides sont indiquées. Les séries de G impliquées dans le rG4 sont encadrées, les G mutés en A dans le mutant sont indiqués en rouge avec un cercle de bordure noire. Les codons de départ alternatifs et le codon de départ de l'uORF sont encadrés et annotés.

**Figure 44** Structure secondaire du 5'UTR complet de BAG-1 WT obtenue par SHAPE

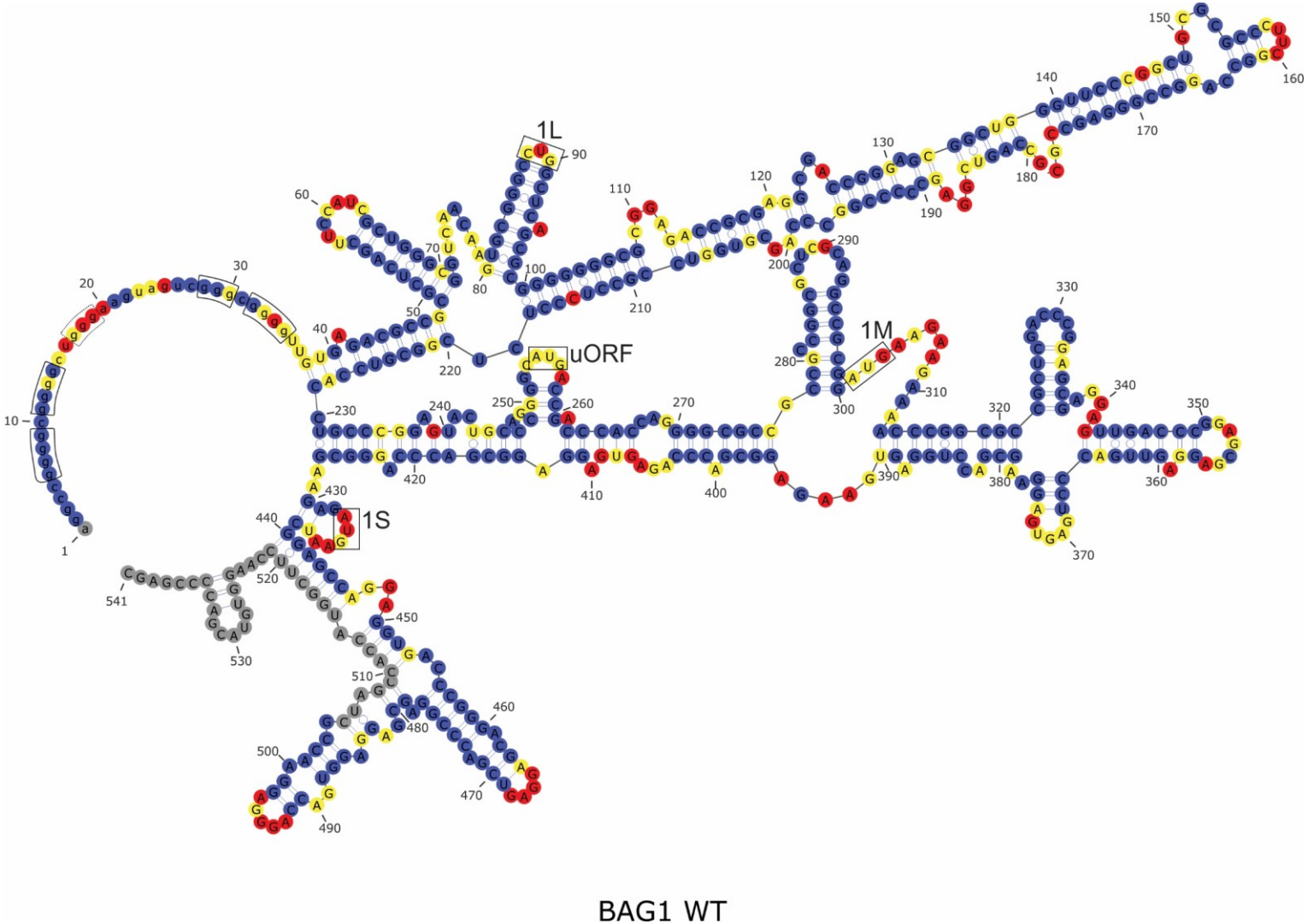


**Figure 45** Structure secondaire du 5'UTR complet de BAG-1 G4mut obtenue par SHAPE